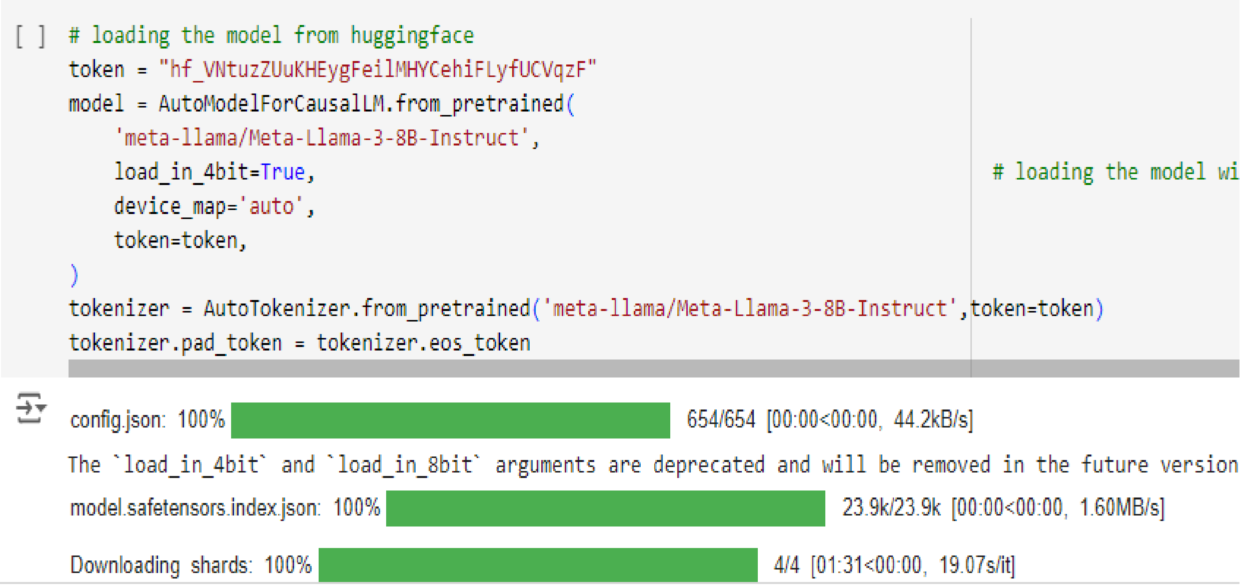
LLM_ASSIGNMENT-1

Mahisha Ramesh
MT23121

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)

1. LLAMA 3.1



```
[ ] # loading the model from huggingface
    token = "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"
    model = AutoModelForCausalLM.from_pretrained(
        'meta-llama/Meta-Llama-3-8B-Instruct',
        load_in_4bit=True,                                                      # loading the model wi
        device_map='auto',
        token=token,
    )
    tokenizer = AutoTokenizer.from_pretrained('meta-llama/Meta-Llama-3-8B-Instruct',token=token)
    tokenizer.pad_token = tokenizer.eos_token
```

```
config.json: 100%  654/654 [00:00<00:00, 44.2kB/s]
The `load_in_4bit` and `load_in_8bit` arguments are deprecated and will be removed in the future version
model.safetensors.index.json: 100%  23.9k/23.9k [00:00<00:00, 1.60MB/s]
Downloading shards: 100%  4/4 [01:31<00:00, 19.07s/it]
```

Output:-

```
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Prompt: group of organisms of the same species that live in the same
Starting from v4.46, the `logits` model output will have the same type as
the model (except at train time, where it will always be FP32)
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Generated Output 1: group of organisms of the same species that live in
the same area
biological community, ecological group, biological population

the individuals are called members or elements and can be described by
their characteristics such as age structure (age distribution), sex ratio
etc.

groups may not always have a fixed boundary. for example:
Generated Output 2: group of organisms of the same species that live in
the same general area (i.e., a population). The individuals within...
a small animal with four legs, typically found on land. It is
characterized by its compact body and limbs adapted for movement.....
the study or science dealing with animals...in their natural
Generated Output 3: group of organisms of the same species that live in
the same area and interbreed, which is called a population. Population
ecology studies how populations respond to their environment,... read
more...
Population Ecology: Definition & Study
https://study.com/academy/population-ecology.html
Learn about what constitutes
The model's responses are not consistent.

Prompt: which is the nearest of the gas giant planets to the sun?
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Generated Output 1: which is the nearest of the gas giant planets to the
sun?...
The correct answer: Jupiter
Jupiter, along with Saturn and Uranus in our solar system are known as Gas
Giant Planets. Among them, Neptune or Venus (not a planet but an
asteroid), among other things it can create confusion about
Generated Output 2: which is the nearest of the gas giant planets to the
sun? A) Jupiter B) Saturn C...  (Read More)

This question was created by a student. You can check out other students'
questions or create your own.

A)
B)
C)


Answer confusion! I'll choose an answer randomly
```

```
Generated Output 3: which is the nearest of the gas giant planets to the
sun?... (More)
The correct answer: Mercury
Mercury, not a Gas Giant Planet! Incorrect question.
Gas giants are massive celestial bodies that orbit very far from their
host star. The solar system's four major gas...
Read more about which is
The model's responses are not consistent.

Prompt: Honey pours more slowly than vinegar because honey has greater
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Generated Output 1: Honey pours more slowly than vinegar because honey has
greater viscosity. Viscosity is the measure of how easily a liquid flows
or resists flow.
In other words, when you pour syrup (honey) and water together into two
separate cups from one bottle cap opened simultaneously at same height
with no air
Generated Output 2: Honey pours more slowly than vinegar because honey has
greater viscosity. Viscosity is a measure of the resistance to flow or
movement between two surfaces in contact with each other.
In this case, it's like trying to pour water through very sticky fingers
while pouring oil on some thin sheets - one flows easily
Generated Output 3: Honey pours more slowly than vinegar because honey has
greater viscosity. Viscosity is the measure of a liquid's resistance to
flowing or its "thickness". Honey, being much thicker and heavier, takes
longer for it to flow through an opening compared with thinner liquids
like water (vinegar). So when you
The model's responses are not consistent.

Prompt: general name given to biochemical polymers
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Generated Output 1: general name given to biochemical polymers
(biodegradable or not) made up of repeating units. They have been used as
biomaterials for many years, such... View More
medical devices and pharmaceutical products using bioplastics.
Bioplastic is a term that has become increasingly
Generated Output 2: general name given to biochemical polymers which
contain repeating units of amino acids, such as proteins and peptides.
Also called polypeptide.
peptides
short chains or sequences of linked amino acid residues; often referred...
simple sugars - monosaccharides (e.g., glucose)
```

```
Generated Output 3: general name given to biochemical polymers such as
proteins, carbohydrates and nucleic acids. These biomolecules are
essential for the functioning of all living organisms...
The term "biomolecule" is used in bioscience research studies that focus
on understanding biological processes through chemical analysis.
Biom
The model's responses are not consistent.

Prompt: scientist who developed atomic theory
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Generated Output 1: scientist who developed atomic theory
What is Atomic Theory? - Definition &
HistoryAtomicTheory.net/what-is-atomic-theory/
https://www.atomictheory.info/scientists/democritus.htmlDemocritos
Demokriteas was a Greek philosopher, one of the
Generated Output 2: scientist who developed atomic theory, discovered
electron and proton
Scientists have made a tremendous amount of progress in understanding the
structure and properties of atoms since Robert Millikan's time. In his
honor we celebrate him as one on Physics Today.
Millikans work was fundamental to our modern
Generated Output 3: scientist who developed atomic theory
What scientist is credited with the development of modern Atomic Theory?
The answer to this question can be found in many different science
textbooks and online resources. In general, it was Robert Millikan's
experiments on beta particles that helped establish a more accurate
understanding
The model's responses are not consistent.

Prompt: amount of space taken up by matter
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Generated Output 1: amount of space taken up by matter in the universe
density is a measure that gives you an idea about how packed or spread-out
things are.
If something has high density, it means there's not much empty space
around what makes us think they're really crowded. If everything else
stays
Generated Output 2: amount of space taken up by matter in a system.
The amount of energy required to heat 1 unit mass (typically grams) or
volume... It is often used as an index for the thermodynamic
properties....
Thermal capacity, also known...: Chemical Engineering Dictionary [home]
Generated Output 3: amount of space taken up by matter and radiation in
the universe
the amount of physical energy produced within a unit time, usually
measured per second (e.g. watts)
a process that occurs when an object is heated or cooled at constant
pressure; it involves compression...
information about how much
The model's responses are not consistent.
```

Explanation of output

Prompt 1: Group of organisms of the same species that live in the same

Consistency: The responses vary in how they define the group of organisms. Some describe it as a "biological community" or "ecological group," while others mention a "population."

Hallucination: The model introduces unrelated concepts like "age structure" and "sex ratio" without directly answering the prompt.

Prompt 2: Which is the nearest of the gas giant planets to the sun?

Consistency: The responses are inconsistent, with one correctly identifying Jupiter, another introducing irrelevant multiple-choice format, and the third incorrectly identifying Mercury, which isn't even a gas giant.

Hallucination: The third response incorrectly labels Mercury as a gas giant, which is a clear hallucination.

Prompt 3: Honey pours more slowly than vinegar because honey has greater

Consistency: All responses agree that honey has greater viscosity, but the explanations vary in detail.

Hallucination: Some responses add unnecessary or incorrect analogies, like "pouring water through sticky fingers," which detracts from the accuracy of the explanation.

Prompt 4: General name given to biochemical polymers

Consistency: The responses are inconsistent, with varied explanations and examples, ranging from bioplastics to polypeptides.

Hallucination: Some responses introduce irrelevant details, such as "bioplastics" and "medical devices," which aren't directly related to biochemical polymers.

Prompt 5: Scientist who developed atomic theory

Consistency: The responses are inconsistent, with varying mentions of John Dalton, Democritus, and Robert Millikan, leading to confusion.

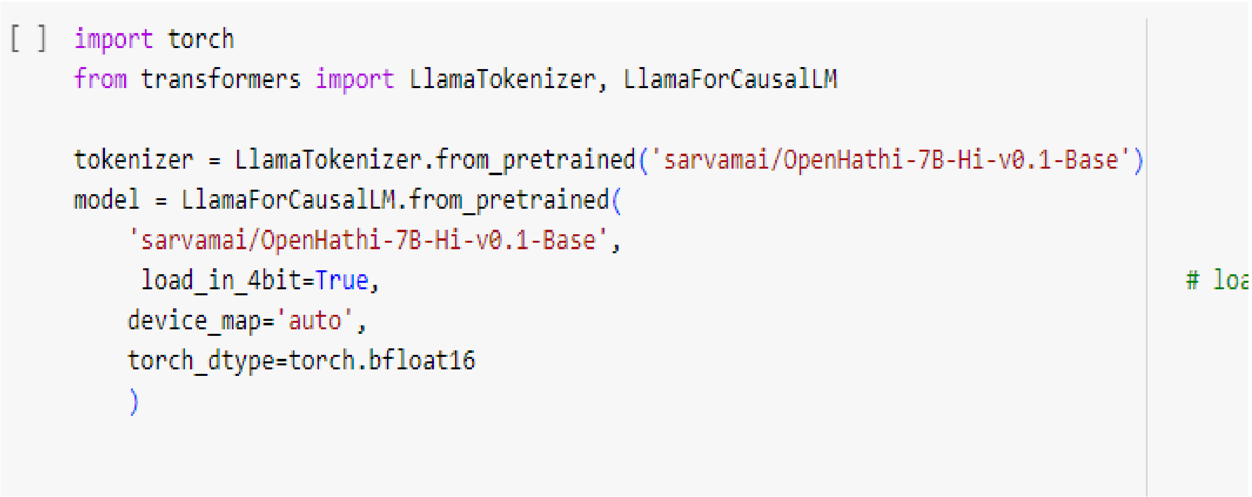
Hallucination: The model incorrectly attributes atomic theory development to Robert Millikan and introduces unrelated links and scientists, showing clear hallucination.

Prompt 6: Amount of space taken up by matter

Consistency: The responses are inconsistent, with varied interpretations of the prompt, ranging from density to thermal capacity.

Hallucination: Some responses introduce irrelevant scientific concepts like "thermal capacity" and "radiation," which aren't directly related to the prompt.

2) OpenHathi-7B



```
import torch
from transformers import LlamaTokenizer, LlamaForCausalLM

tokenizer = LlamaTokenizer.from_pretrained('sarvamai/OpenHathi-7B-Hi-v0.1-Base')
model = LlamaForCausalLM.from_pretrained(
    'sarvamai/OpenHathi-7B-Hi-v0.1-Base',
     load_in_4bit=True,                                          # loa
    device_map='auto',
    torch_dtype=torch.bfloat16
    )
```

Output:-

```
Prompt 1: the 2004 boxing day tsunami killed people all around this ocean
basin
```

```
Output 1-1: the 2004 boxing day tsunami killed people all around this
ocean basin. संतुलित महासागर बेसिनों में, सुनामी का खतरा कम होता है।
The Indian Ocean is the most important source of tsunam
Output 1-2: the 2004 boxing day tsunami killed people all around this
ocean basin. संतुलित महासागर बेसिनों में, सुनामी का खतरा कम होता है।
The Indian Ocean is the most important source of tsunam
Output 1-3: the 2004 boxing day tsunami killed people all around this
ocean basin. संतुलित महासागर बेसिनों में, सुनामी का खतरा कम होता है।
The Indian Ocean is the most seismically active ocean bas
Consistent: No

--------------------------------------------------

Prompt 2: Which state of matter has particles with the least energy?
Output 2-1: Which state of matter has particles with the least energy?
 संतुलित अवस्था में गैसें।
What is the difference between a gas and a liquid?
तरल पदार्थ में कणों की गति कम होती है।
What is the difference between
Output 2-2: Which state of matter has particles with the least energy?
 संतुलित अवस्था में गैसें।
What is the difference between a gas and a liquid?
तरल पदार्थ में कणों की गति कम होती है।
What is the difference between
Output 2-3: Which state of matter has particles with the least energy?
 संतुलित अवस्था में गैसें।
What is the difference between a gas and a liquid?
तरल पदार्थ में कणों की गति कम होती है।
What is the difference between
Consistent: Yes

--------------------------------------------------

Prompt 3: largest and most common parasitic worm in humans
Output 3-1: largest and most common parasitic worm in humans. संतुलित,
100-150 मिमी लंबा, 1-2 मिमी चौड़ा, और 100-200 मिमी लंबा, 1-2
Output 3-2: largest and most common parasitic worm in humans. संतुलित,
100-150 मिमी लंबा, 1-2 मिमी चौड़ा, और 100-200 मिमी लंबा, 1-2
Output 3-3: largest and most common parasitic worm in humans. संतुलित,
100-150 मिमी लंबा, 1-2 मिमी चौड़ा, और 100-200 मिमी लंबा, 1-2
Consistent: Yes

--------------------------------------------------

Prompt 4: Underground water is heated by
```

```
Output 4-1: Underground water is heated by the Earth's core and then rises
to the surface. संतुलित जल चक्र के हिस्से के रूप में, यह फिर सतह के पानी में वापस चला
जाता है।

2. The water cycle is a
Output 4-2: Underground water is heated by the Earth's core and then rises
to the surface. संतुलित जल चक्र के हिस्से के रूप में, यह फिर सतह के पानी में वापस चला
जाता है।

2. The water cycle is a
Output 4-3: Underground water is heated by the Earth's core and then rises
to the surface. संतुलित जल चक्र के हिस्से के रूप में, यह फिर सतह के पानी में वापस चला
जाता है।

2. The water cycle is a
Consistent: Yes

--------------------------------------------------

Prompt 5: philosopher who thought the idea of the atom was ridiculous
Output 5-1: philosopher who thought the idea of the atom was ridiculous.
संतुलित, और एक ऐसे युग में जब अधिकांश दार्शनिकों ने सोचा था कि परमाणु एक वास्तविक वस्तु
थी, डेसकार्टेस ने सोचा कि यह
Output 5-2: philosopher who thought the idea of the atom was ridiculous.
संतुलित, और एक ऐसे युग में जब अधिकांश दार्शनिकों ने सोचा था कि परमाणु एक वास्तविक वस्तु
थी, डेसकार्टेस ने सोचा कि यह
Output 5-3: philosopher who thought the idea of the atom was ridiculous.
संतुलित, और एक ऐसे युग में जब अधिकांश दार्शनिकों ने सोचा था कि परमाणु एक वास्तविक वस्तु
थी, डेसकार्टेस ने सोचा कि यह
Consistent: Yes

--------------------------------------------------

Prompt 6: circular or elliptical path around an object
Output 6-1: circular or elliptical path around an object. संतुलित और कैंद्रित।

10. "The Earth's rotation is the reason for day and night."

11. "सूर्य पृथ्वी के चारों ओर घूमता है।
Output 6-2: circular or elliptical path around an object. संतुलित और कैंद्रित।

10. "The Earth's rotation is the reason for day and night."

11. "सूर्य पृथ्वी के चारों ओर घूमता है।
Output 6-3: circular or elliptical path around an object. संतुलित और कैंद्रित।

10. "The Earth's rotation is the reason for day and night."
```

```
11. "सूर्य का प्रकाश पृथ्वी पर आता है।"
Consistent: No

--------------------------------------------------
```

Explanation of output:-

Prompt 1: The 2004 Boxing Day tsunami killed people all around this ocean basin

Consistency: The responses are consistent in structure, but all include a non-English phrase ("संतुलित महासागर बेसिनों में, सुनामी का खतरा कम होता है") and incomplete sentences, indicating possible hallucination.

Hallucination: The phrase in Hindi is unrelated to the prompt, and the sentence about "tsunami risk being lower" is incorrect in this context. The mention of "The Indian Ocean is the most important source of tsunam" is cut off and suggests incomplete thought processing.

Prompt 2: Which state of matter has particles with the least energy?

Consistency: The responses are consistent, repeating the same structure and phrasing. Hallucination: The repeated Hindi phrase ("संतुलित अवस्था में गैसें") incorrectly suggests gases as having the least energy, which contradicts the scientific fact that solids have the least energy.

Prompt 3: Largest and most common parasitic worm in humans.

Consistency: The responses are consistent, repeating the same structure and phrasing. Hallucination: The responses provide an incomplete description ("संतुलित, 100-150 मिमी लंबा..."), and the information presented is not specific enough to identify the parasitic worm, leading to potential hallucination.

Prompt 4: Underground water is heated by...

Consistency: The responses are consistent in structure and content.

Hallucination: The sentence in Hindi ("संतुलित जल चक्र के हिस्से के रूप में") is unrelated to the scientific explanation of geothermal heating, indicating hallucination.

Prompt 5: Philosopher who thought the idea of the atom was ridiculous.

Consistency: The responses are consistent, repeating the same structure and phrasing.

Hallucination: The mention of "Descartes" in the context of atomic theory is misleading, as Descartes did not specifically oppose the idea of atoms. This suggests the model is fabricating information.

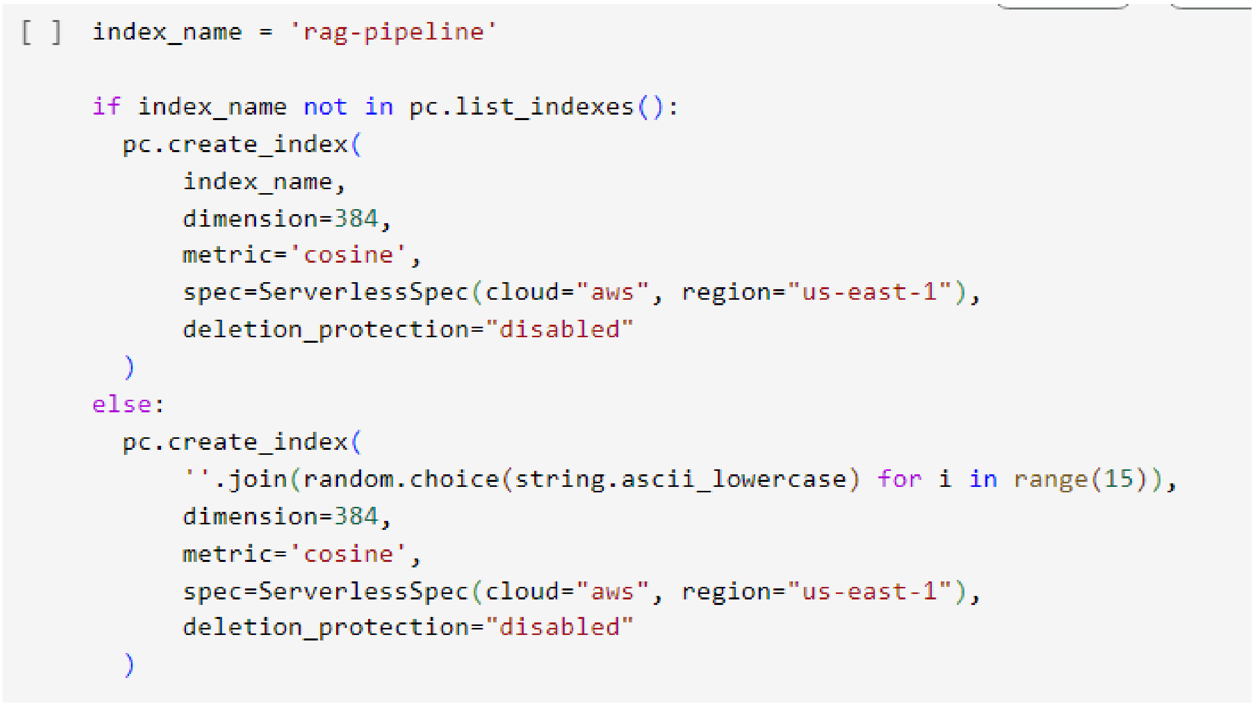
Prompt 6: Circular or elliptical path around an object.

Consistency: The responses are generally consistent, but the last output slightly differs.

Hallucination: The mention of unrelated concepts like the Earth's rotation and the Sun's light ("सूर्य का प्रकाश पृथ्वी पर आता है") is irrelevant to the prompt, indicating hallucination.

Task 2 (6 Marks) :

○ **Use Retrieval-Augmented Generation (RAG) techniques to minimize or solve all the hallucinations identified in the previous step. (6 Marks)**



```
index_name = 'rag-pipeline'

if index_name not in pc.list_indexes():
  pc.create_index(
      index_name,
      dimension=384,
      metric='cosine',
      spec=ServerlessSpec(cloud="aws", region="us-east-1"),
      deletion_protection="disabled"
  )
else:
  pc.create_index(
      ''.join(random.choice(string.ascii_lowercase) for i in range(15)),
      dimension=384,
      metric='cosine',
      spec=ServerlessSpec(cloud="aws", region="us-east-1"),
      deletion_protection="disabled"
  )
```

LLAMA 3.1-INSTRUCT With RAG

```
{'query': 'group of organisms of the same species that live in the same',
'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\n\n\nQuestion: group of organisms of the same species
that live in the same\nHelpful Answer: population\n\nExplanation: A population
is a group of organisms of the same species that live in the same"}

{'query': 'which is the nearest of the gas giant planets to the sun?',
'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\n\n\nQuestion: which is the nearest of the gas giant
planets to the sun?\nHelpful Answer: Jupiter is the nearest gas giant to the
Sun, with an average distance of about 778.3"}

{'query': 'Honey pours more slowly than vinegar because honey has greater',
'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\n\n\nQuestion: Honey pours more slowly than vinegar
because honey has greater\nHelpful Answer: viscosity\nHelpful Answer:
density\nHelpful Answer: surface tension\nHelpful Answer: gravity"}

{'query': 'general name given to biochemical polymers', '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\n\n\nQuestion: general name given to biochemical polymers\nHelpful
Answer: Polysaccharides\n\nExplanation: Biochemical polymers, such as
cellulose, glycogen"}

{'query': 'scientist who developed atomic theory', '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\n\n\nQuestion: scientist who developed atomic theory\nHelpful
Answer: John Dalton\n\nAdditional Context: John Dalton was an English chemist
and physicist who is widely credited with"}

{'query': 'amount of space taken up by matter', '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\n\n\nQuestion: amount of space taken up by matter\nHelpful Answer:
volume\n\nHelpful Answer: mass\n\nHelpful Answer: density\n\nHelpful Answer:
surface area"}
```

openHaithi with RAG

```
{'query': 'the 2004 boxing day tsunami killed people all around this ocean
basin', '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\n\n\nQuestion: the 2004 boxing day
tsunami killed people all around this ocean basin\nHelpful Answer:
220,000\nसंदर्भः 1. The"}
```

```
{'query': 'Which state of matter has particles with the least energy?',
'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\n\n\nQuestion: Which state of matter has
particles with the least energy?\nHelpful Answer:\n---\nचरण 1: प्रश्न के संदर्भ
को समझें\nThe question is"}

{'query': 'largest and most common parasitic worm in humans', '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\n\n\nQuestion: largest and most common parasitic worm in
humans\nHelpful Answer: tapeworm\n\n\n\nप्रश्नः सबसे बड़ा और सबसे आम पर"}

{'query': 'Underground water is heated by', '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\n\n\nQuestion: Underground water is heated by\nHelpful Answer:
Underground water is heated by the heat of the earth's core.\n\n\n\n"}

{'query': 'philosopher who thought the idea of the atom was ridiculous',
'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\n\n\nQuestion: philosopher who thought the
idea of the atom was ridiculous\nHelpful Answer:\n---\nThe philosopher who
thought the idea of the atom was ridiculous was Dem"}

{'query': 'circular or elliptical path around an object', '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\n\n\nQuestion: circular or elliptical path around an
object\nHelpful Answer: elliptical\n\n\n\nप्रश्नः एक वस्तु के चारों ओर गोलाकार"}
```

How RAG improves the performance of the models ?

Informed Responses: RAG improves the accuracy of the generated outputs because the model is not entirely dependent on its internal knowledge (which can become outdated or insufficient). It supplements generation with up-to-date and relevant information from the external corpus.
Reduction of Hallucination: Purely generative models can sometimes "hallucinate" facts, generating incorrect or fabricated details. RAG mitigates this by grounding the generation in actual retrieved documents, which improves factual correctness.

PART 2 - Probing

Objective: The task is to explore how well Large Language Models (LLMs) encode information
about various topics or entities at different layers. You will use probing techniques to analyze the
model's ability to retain and predict specific information based on a dataset

Steps:

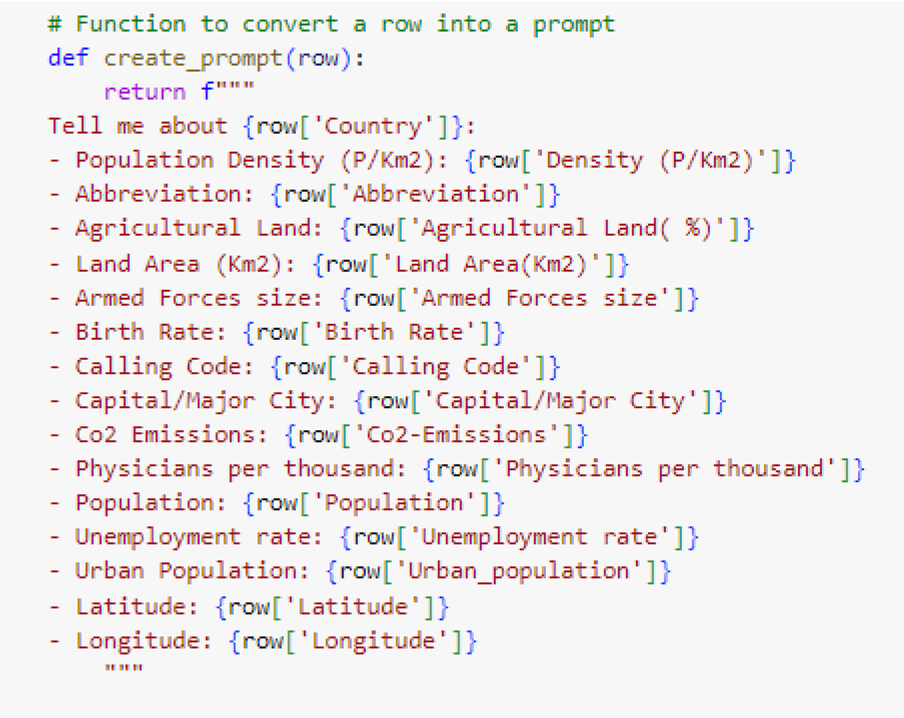
1. Select a Dataset:

○ Choose a dataset that contains structured information about a specific topic or entities (e.g., historical figures, geographical locations, scientific concepts, people).

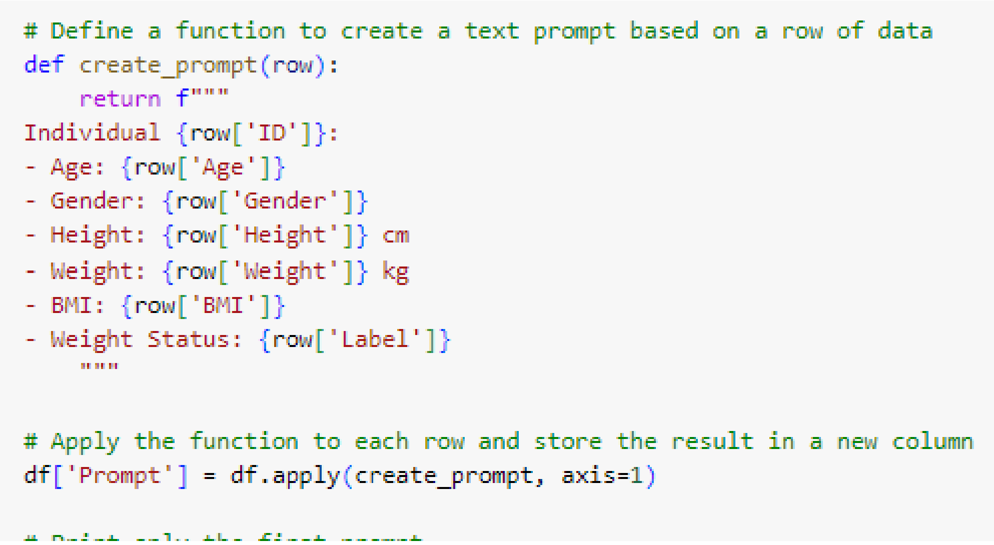
○ Ensure your dataset has several fields that can be predicted (e.g.,, population, notable achievements, dates, any number, any class).

2. Design a Prompt:

○ Create prompts that query the LLM about the entities or topics in the dataset

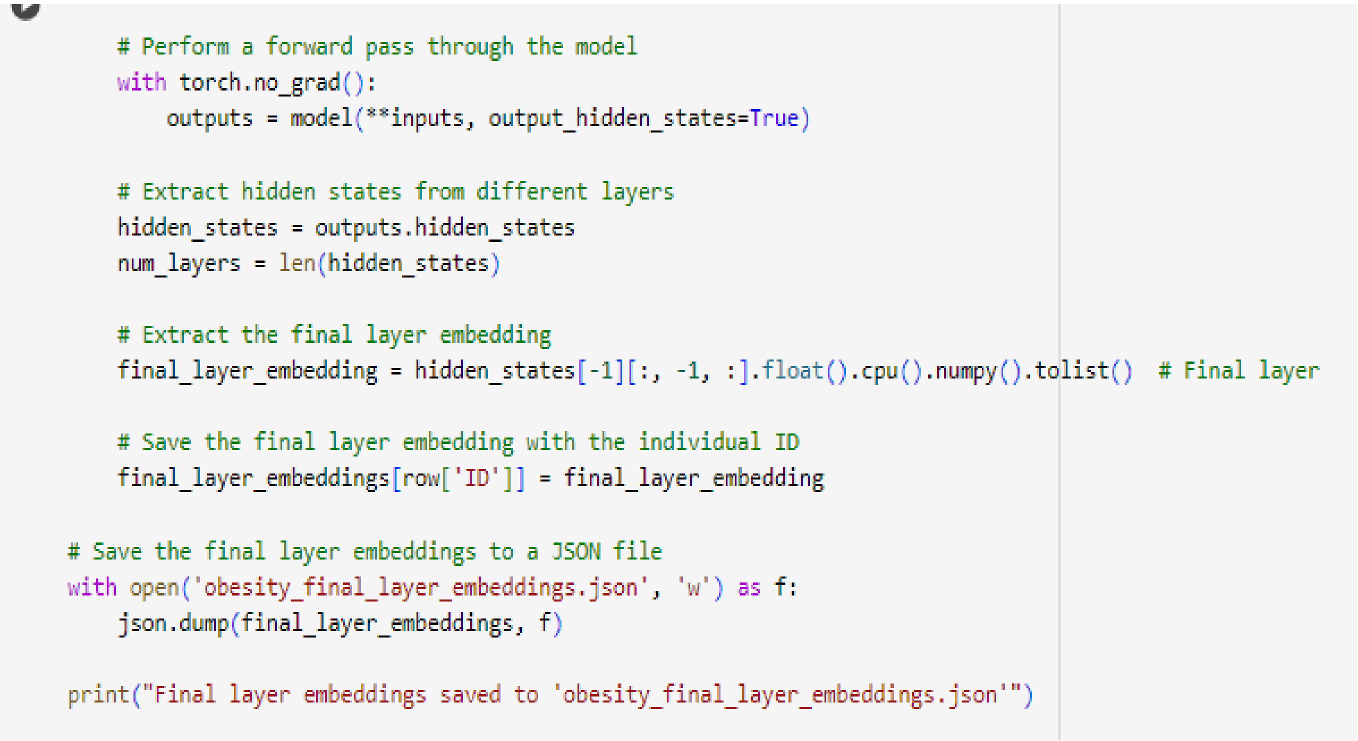


```python
# Function to convert a row into a prompt
def create_prompt(row):
    return f"""
Tell me about {row['Country']}:
- Population Density (P/Km2): {row['Density (P/Km2)']}
- Abbreviation: {row['Abbreviation']}
- Agricultural Land: {row['Agricultural Land( %)']}
- Land Area (Km2): {row['Land Area(Km2)']}
- Armed Forces size: {row['Armed Forces size']}
- Birth Rate: {row['Birth Rate']}
- Calling Code: {row['Calling Code']}
- Capital/Major City: {row['Capital/Major City']}
- Co2 Emissions: {row['Co2-Emissions']}
- Physicians per thousand: {row['Physicians per thousand']}
- Population: {row['Population']}
- Unemployment rate: {row['Unemployment rate']}
- Urban Population: {row['Urban_population']}
- Latitude: {row['Latitude']}
- Longitude: {row['Longitude']}
    """
```



```
# Define a function to create a text prompt based on a row of data
def create_prompt(row):
    return f"""
Individual {row['ID']}:
- Age: {row['Age']}
- Gender: {row['Gender']}
- Height: {row['Height']} cm
- Weight: {row['Weight']} kg
- BMI: {row['BMI']}
- Weight Status: {row['Label']}
    """

# Apply the function to each row and store the result in a new column
df['Prompt'] = df.apply(create_prompt, axis=1)

# Print only the first prompt
```

3) Extract Embeddings:
○ Use your designed prompts to perform a forward pass through an LLM (Use LLAMA 3) and extract the embedding of the final token.



```
        # Perform a forward pass through the model
        with torch.no_grad():
            outputs = model(**inputs, output_hidden_states=True)

        # Extract hidden states from different layers
        hidden_states = outputs.hidden_states
        num_layers = len(hidden_states)

        # Extract the final layer embedding
        final_layer_embedding = hidden_states[-1][:, -1, :].float().cpu().numpy().tolist()  # Final layer

        # Save the final layer embedding with the individual ID
        final_layer_embeddings[row['ID']] = final_layer_embedding

    # Save the final layer embeddings to a JSON file
    with open('obesity_final_layer_embeddings.json', 'w') as f:
        json.dump(final_layer_embeddings, f)

    print("Final layer embeddings saved to 'obesity_final_layer_embeddings.json'")
```

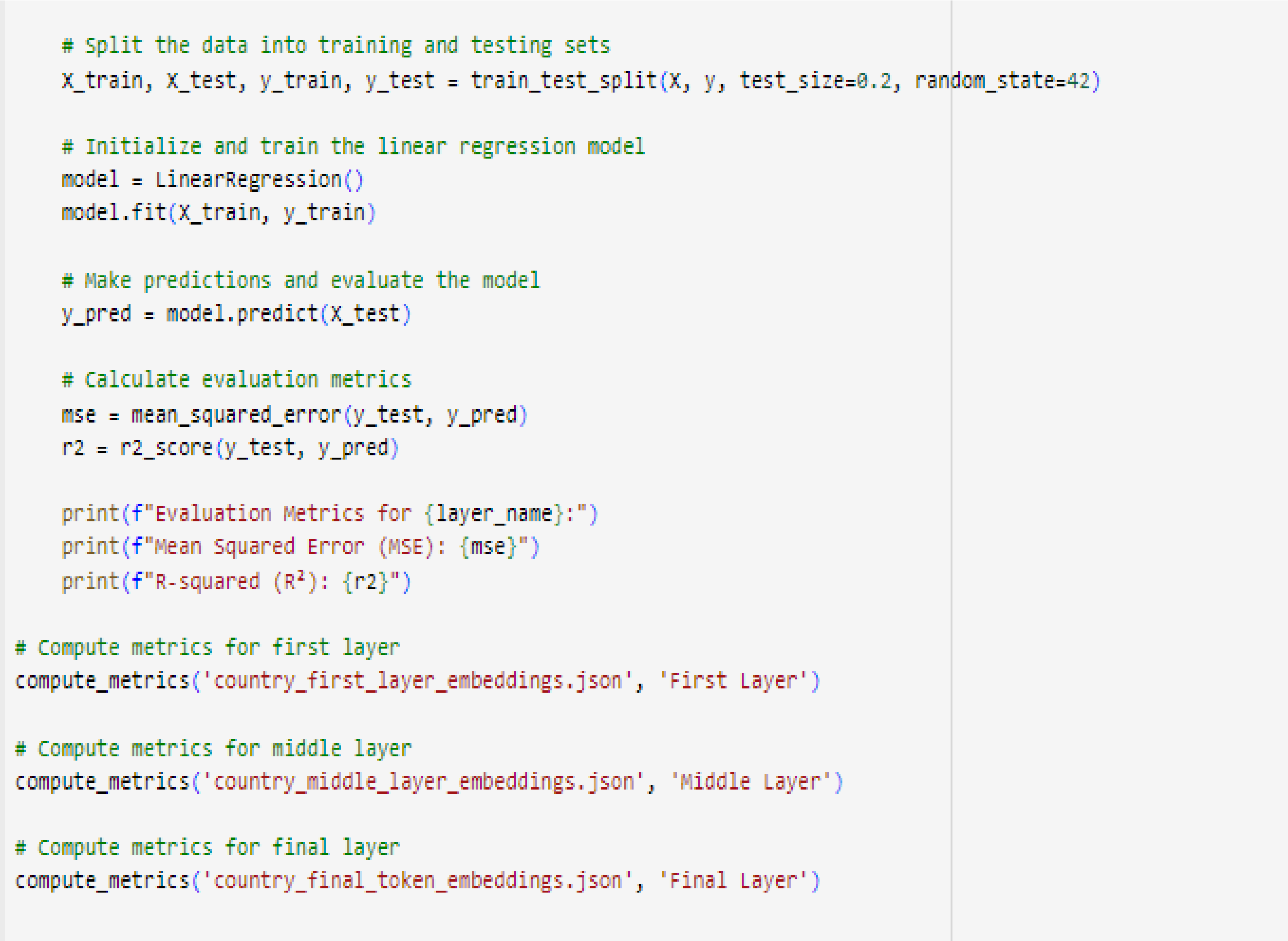
4) Set Up a Linear Regression and Classification model:

○ Use the extracted token embeddings as inputs for a linear regression model and a classification model (Regressor and classifier are just like a head that can extract info from embedding). You will aim to predict several fields from your chosen dataset, such as:

■ A numeric field using regression: (e.g., birth year, population, discovery date, number of mentions, page views, citations, etc).

■ A class variable:(e.g: Gender, profession, etc).

Regressor based on probing

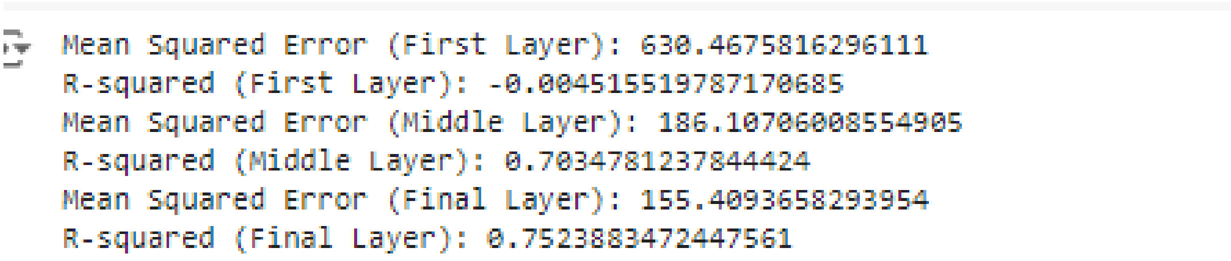


```
    # Split the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Initialize and train the linear regression model
    model = LinearRegression()
    model.fit(X_train, y_train)

    # Make predictions and evaluate the model
    y_pred = model.predict(X_test)

    # Calculate evaluation metrics
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)

    print(f"Evaluation Metrics for {layer_name}:")
    print(f"Mean Squared Error (MSE): {mse}")
    print(f"R-squared (R²): {r2}")

# Compute metrics for first layer
compute_metrics('country_first_layer_embeddings.json', 'First Layer')

# Compute metrics for middle layer
compute_metrics('country_middle_layer_embeddings.json', 'Middle Layer')

# Compute metrics for final layer
compute_metrics('country_final_token_embeddings.json', 'Final Layer')
```

Birth Rate values

```
Evaluation Metrics for First Layer:
Mean Squared Error (MSE): 111.23680819146897
R-squared (R²): -0.0012445280696453676
Evaluation Metrics for Middle Layer:
Mean Squared Error (MSE): 19.144070985938228
R-squared (R²): 0.8276838698333171
Evaluation Metrics for Final Layer:
Mean Squared Error (MSE): 13.28294735052522
R-squared (R²): 0.8804399499807786
```

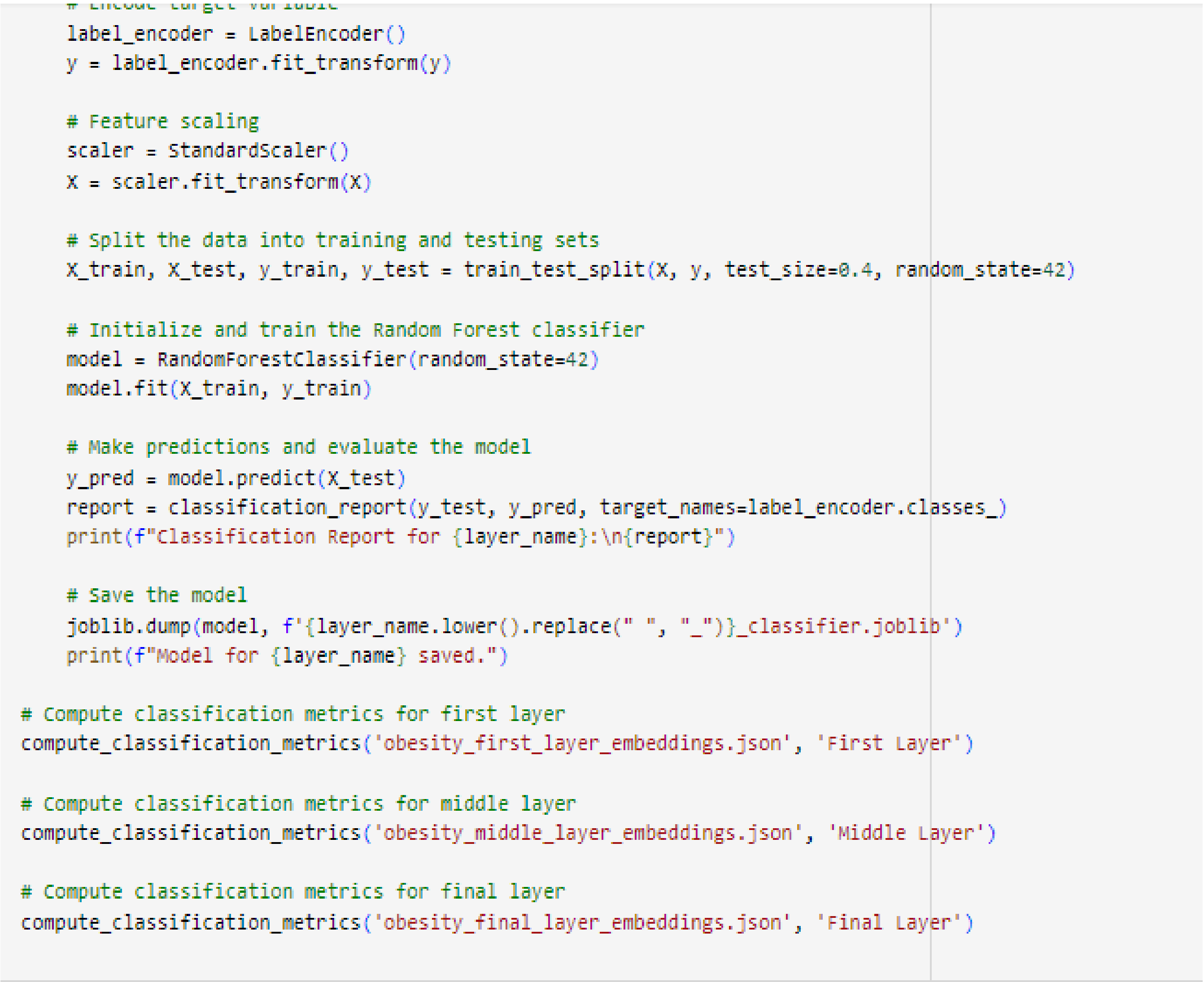
Latitude:-



```
Mean Squared Error (First Layer): 630.4675816296111
R-squared (First Layer): -0.004515519787170685
Mean Squared Error (Middle Layer): 186.10706008554905
R-squared (Middle Layer): 0.7034781237844424
Mean Squared Error (Final Layer): 155.4093658293954
R-squared (Final Layer): 0.7523883472447561
```

ouble-click (or enter) to edit

Classification based probing



```
    # Encode target variable
    label_encoder = LabelEncoder()
    y = label_encoder.fit_transform(y)

    # Feature scaling
    scaler = StandardScaler()
    X = scaler.fit_transform(X)

    # Split the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)

    # Initialize and train the Random Forest classifier
    model = RandomForestClassifier(random_state=42)
    model.fit(X_train, y_train)

    # Make predictions and evaluate the model
    y_pred = model.predict(X_test)
    report = classification_report(y_test, y_pred, target_names=label_encoder.classes_)
    print(f"Classification Report for {layer_name}:\n{report}")

    # Save the model
    joblib.dump(model, f'{layer_name.lower().replace(" ", "_")}_classifier.joblib')
    print(f"Model for {layer_name} saved.")

# Compute classification metrics for first layer
compute_classification_metrics('obesity_first_layer_embeddings.json', 'First Layer')

# Compute classification metrics for middle layer
compute_classification_metrics('obesity_middle_layer_embeddings.json', 'Middle Layer')

# Compute classification metrics for final layer
compute_classification_metrics('obesity_final_layer_embeddings.json', 'Final Layer')
```

```
Classification Report for First Layer:
               precision    recall  f1-score   support

Normal Weight       0.00      0.00      0.00         8
        Obese       0.00      0.00      0.00         7
   Overweight       0.00      0.00      0.00        12
  Underweight       0.39      1.00      0.56        17

     accuracy                           0.39        44
    macro avg       0.10      0.25      0.14        44
 weighted avg       0.15      0.39      0.22        44

Model for First Layer saved.
Classification Report for Middle Layer:
```

```
           precision    recall  f1-score   support

Normal Weight       1.00      1.00      1.00         8
      Obese       1.00      1.00      1.00         7
  Overweight       1.00      1.00      1.00        12
 Underweight       1.00      1.00      1.00        17

    accuracy                           1.00        44
   macro avg       1.00      1.00      1.00        44
weighted avg       1.00      1.00      1.00        44

Model for Middle Layer saved.
Classification Report for Final Layer:
           precision    recall  f1-score   support

Normal Weight       1.00      1.00      1.00         8
      Obese       1.00      1.00      1.00         7
  Overweight       1.00      1.00      1.00        12
 Underweight       1.00      1.00      1.00        17

    accuracy                           1.00        44
   macro avg       1.00      1.00      1.00        44
weighted avg       1.00      1.00      1.00        44

Model for Final Layer saved.
```

---

6. Discussion:
○ Reflect on your findings. What do the results indicate about the LLM's ability to
encode the information in your dataset?
○ Discuss any patterns or anomalies you observe, such as differences in performance between various models or layers.

For Regressor based Probing

- **First Layer:** The model performs poorly, as the embeddings contain little useful information for the task. This is typical because early layers of a model are usually responsible for capturing low-level features.

- **Middle Layer:** There is a **huge improvement** in both MSE and R^2, suggesting that the middle layers capture more useful, abstracted features that are important for the task. The model is able to explain much more of the target variance at this stage.
- **Final Layer:** Performance further improves, showing that the final layer embeddings are the most informative for this task. They contain highly processed and task-specific information, leading to the best predictive performance.

The model's ability to predict the target improves as you move deeper into the layers. The final layer provides the most refined information, resulting in the best performance. This indicates that the LLM gradually encodes more task-relevant information as you move towards the final layers.

Classification based probing

1. **First Layer:**
   - **Poor performance:** The model struggles to classify correctly for most classes.
   - **Why?:** The first layer embeddings likely represent lower-level, more general features that are not specific enough for accurate classification. They may capture surface-level patterns but not the deeper relationships needed for the task.
2. **Middle Layer:**
   - **Perfect performance:** All metrics are perfect (1.00), which shows the middle layer captures rich, relevant features for classification.
   - Middle layers in models like LLAMA-3 often capture abstract patterns and task-specific representations, making them highly effective for tasks like classification.
3. **Final Layer:**
   - **Also perfect performance:** Similar to the middle layer, but no significant improvement.
   - Final layers fine-tune the embeddings to task-specific goals, but in this case, the middle layer embeddings were already optimal. The final layer might be refining these representations slightly without impacting performance further.