

CSE665: Large Language Models

Assignment 1

PART 1 - Exploring Large Language Models

Examples of self-consistency questions:

- You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?
- Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?
- A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll?

Examples of fact-checking questions:

- Which Indian won two bronze medals at the 2024 Paris Olympics?
- Who won the Women's Singles US Open 2024?
- Which company acquired Splunk?

OpenHathi Self Consistency

Question: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?

Response 1: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?

A) 40 डिग्री। B)...92° C) .86 ° D).....75 ° E) ..यहाँ से बाहर निकलें

Response 2: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?

A) 20 डिग्री। B) 48 degree C)...76° D.)...98 ° E.)..... नहीं, कोई उत्तर सही या गलत है: A is correct because one of a watch's hands points to every five minutes on an analogue dial . घड़ी के घंटे और मिनट हाथ एक दूसरे से अलग होते हैं ताकि आप देख सकें कि वे हर पांच सेकंड में कैसे आगे बढ़ते रहते हैं (जैसे जब हम घंटों को देखते हुए कहते थे "घंटे की ओर देखो" तो यह कहने का मतलब था जैसे ही आपने कहा शब्द समाप्त किया वह अभी भी अपने स्थान पर खड़ा होगा), लेकिन केवल तभी आपको पता चलता अगर आपके पास घड़ियां हों जो हमेशा सटीक होती थीं क्योंकि उन्होंने उस

Analysis: The OpenHathi model produces inconsistent responses because it seems to mix languages and provide answers that don't directly address the question in a coherent or logical way.

Response 1 includes random numbers like 40°, 92°, and 75° without proper reasoning, mixing these with Hindi text. It also includes irrelevant phrases such as "यहाँ से बाहर निकलें" (meaning "exit from here"), which has no relation to the question.

Response 2 repeats similar issues, presenting random angles like 20°, 48°, 76°, and 98°, with unnecessary and confusing explanations in both English and Hindi that do not clearly address how to find the correct angle between the clock hands.

Question: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

Response 1: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

A:

Response 2: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

A:

Analysis: The OpenHathi model sometimes fails to provide an answer, as shown in both responses where the output is blank (just "A:"). This issue may arise from the model's inability to process complex relational logic or interpret the family tree correctly, leading to no response at all.

Question: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

Response 1: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

A:

मार्क सारा से संबंधित है क्योंकि वह लिसा और पॉल का बेटा है।

Response 2: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

A:

मार्क सारा से संबंधित है।

Analysis: In one response, the model provides extra but incomplete details about Mark being Lisa and Paul's son, while the other response simply states that "Mark is related to Sarah" without clarifying how. Neither response directly answers that Mark is Sarah's **cousin**, leading to inconsistent and unclear answers.

Question: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll?

Response 1: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll?

--- 1) नीचे। Since there's no other information given, we can only assume that it would be downhill because gravity pulls things towards Earth (or in this case toward to bottom).

Response 2: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll?

A: ऊपर की ओर।

Analysis: The OpenHathi model gives inconsistent responses because it doesn't correctly interpret key details of the question. For example, it fails to recognize that a **rooster can't lay eggs**, leading to irrelevant answers like "downhill" in one response and "upward" in another.

OpenHathi Fact Checking

The responses provided by the OpenHathi model exhibit clear issues with fact-checking and coherence.

Question: Which Indian won two bronze medals at the 2024 Paris Olympics?

Response: Which Indian won two bronze medals at the 2024 Paris Olympics?
nobody knows

Analysis: This is incorrect and unhelpful. The model should either mention the correct athlete or state that this event hasn't happened yet, depending on the context.

Question: Who won the Women's Singles US Open 2024?

Response: Who won the Women's Singles US Open 2024?
Hopefully, it will be a great match!

Analysis: This response doesn't answer the question and instead provides an irrelevant statement. It suggests the model cannot retrieve the correct information or understand that it should be addressing a factual inquiry.

Question: Which company acquired Splunk?

Response: Which company acquired Splunk?

संतुलित निवेश। 1. What is the name of this investment strategy, and what are its main characteristics ? समझाएँ कि यह एक विशिष्ट प्रकार के पोर्टफोलियो से कैसे संबंधित है (जैसे विविधीकरण या लीवरेज्ड खरीद)। - The balance sheet must be in a positive cash flow position .- परिसंपत्तियों और देनदारियों को उनके वर्तमान मूल्य पर निर्धारित किया जाना चाहिए न कि उनकी मूल लागतों में.- एक व्यक्तिगत निवेशक द्वारा अपने स्वयं के धन का उपयोग करके इस रणनीति की स्थापना करना संभव हो सकता है; हालांकि लाभार्जित भुगतान प्राप्त करने वाले शेयरों जैसे अन्य वित्तीय साधनों तक पहुँचने वाली किसी भी कंपनी/संस्थान से नहीं खरीदना होगा। This can also include

Analysis: The response is incoherent and irrelevant to the question, as it diverges into a discussion about investment strategies in Hindi, with no mention of Splunk. This is a clear hallucination, where the model produces text that is not connected to the input question.

LLAMA3 Self Consistency

Question: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?

Response 1: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand? (The answer is an integer.)

Final Answer: The final answer is 165° . I hope it helps.

I'm not sure if this problem can be solved without knowing that a full circle has 360° , but here's my attempt:

Let

θ_1, θ_2 and $f(x)$ denote angles in degrees between hour hand and minute hands respectively at x o'clock. At x o'clock, $f(x) = \left\{ \begin{array}{l} \text{Mathop *0 :mathend} \end{array} \right\}$ when $x=12$ $f(12) = \frac{x}{5} \cdot 30$. At any other time t , $\text{t:}\text{textbf{n}}$

Response 2: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand? (Assume that it's not exactly quarter past three.)

The hands of a watch move 2π radians per twelve hours, or

$\frac{\text{degrees}}{(12)(360)} = \boxed{\frac{5}{6}}$. So for every five minutes they cover one degree.

Final Answer: The final answer is $\left(1 - \frac{1}{2}\right)$. I hope this will help you.

I'm guessing there may be an issue with my solution because when checking on other times like half-past four where both lines meet each other, we get no error

Please correct me if i am wrong

Another thing to note here would also include whether these two angles form linear pairs which means their sum equals 90

Analysis:

- Response 1 states the final answer is " 165° ," which is incorrect. At 3:15, the angle between the hour and minute hand is actually 7.5° , not 165° .
- Response 2 mixes radians and degrees and provides incorrect equations. The final answer is " $(1 - \frac{1}{2})(1 - \frac{1}{2})$," which is incomplete and nonsensical.

Question: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

Response 1: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah? A) Grandson B) Nephew C) Cousin D) Brother

We can start by identifying each person in this family tree:

- Father: Michael (Sarah's dad)
- * Mother (? unknown for now!)
 - o Child 1 : Sarah
 - + Parent = Michael

George -> Children:

- * Daughter:???
- + Son:???? → This means there are still some people missing!

Lisa & her spouse? Paul Mark?

Let us fill out more details about these characters.

The information given states that both parents will be mentioned.

So we'll find them first! The question says "How [is] MARK Related To SARAH?" So let's look at what happens when you follow back from

Response 2: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah? A) Brother B) Cousin C) Grandson D) Nephew

Sarah's grandpa (George)

has 2 kids:

- Mike - her dad

Lisa

Mike = Dad

Mark -> nephew or cousin?

Answer choice:

B). That would be correct! So there you go!

Final Answer: The final answer is indeed "Cousins". I hope it helps clarify things for anyone struggling with this question.

Analysis:

- Response 1 does not provide a clear final answer and introduces unnecessary confusion by discussing missing information and asking questions about additional details. It fails to deliver a coherent or consistent resolution to the question about Mark's relationship to Sarah.
- Response 2 ultimately provides the correct relationship ("Cousin") but does so with a lack of detailed explanation.

The first response is unclear and incomplete, while the second response correctly identifies the relationship as "Cousin." This discrepancy between the two responses illustrates inconsistency in the model's answers, as it provides different and conflicting information about the same question.

Question: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll?

Response 1: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll? (Assume it's not broken and doesn't have any cracks.) The answer, surprisingly enough, does depend slightly upon whether you're thinking like a physicist or just going with your gut.

From my perspective as someone who likes to think about these sorts of things:

1) In physics terms: Since there are no external forces acting directly onto this "egg" from outside influences such... [Read More](#)

I'll admit that I was initially skeptical when reading...

0 comments so far

Post Your Comment Cancel Reply No HTML allowed in replies! Please use Markdown code for formatting if needed.

* * _italic_ *

Inline Code `code`

Link [link](http://www.example.com)

Blockquote > quote text here <

Response 2: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll? (Assume it's smooth and there are no obstacles.)

The answer, surprisingly enough, does not depend solely upon gravity! It depends also...

...on your perspective!

From any point in space outside the Earth or inside another planet with negligible air resistance...

****There isn't one direction for eggs to fall; they'll always follow geodesics**, which means that their path won't be affected by our intuition about up-down directions.**

In other words: ****the world doesn't care whether you're standing upright like humans do****

Now let's get back down-to-earth!

When we're talking within this atmosphere – where friction plays its part...

Gravity pulls everything towards itself.

On flat surfaces:

- An object placed at rest stays put

(No

Analysis:

- Response 1 discusses the physics of an egg rolling on a sloped roof, which is irrelevant because roosters do not lay eggs.
- Response 2 also explores complex concepts and perspectives, which are irrelevant to the core issue.

The discrepancy between the two responses illustrates inconsistency in the model's answers, as it provides different and conflicting information about the same question.

LLAMA3 Fact Checking

Question: Which Indian won two bronze medals at the 2024 Paris Olympics?

Response: Which Indian won two bronze medals at the 2024 Paris Olympics? A) Neeraj Chopra B) PV Sindhu C) Pranav Sood D) Akash Singh

The correct answer is not mentioned in the options. The winner of the Bronze medal (2 times) was Sharath Kamal, an India table tennis player.

Sharath Kamal became only the second athlete to win multiple Olympic bronzes after winning a silver and one more bronze with his partner Ayhik Gogoi.

Note: This question refers to recent event which happened recently so it's possible that you may have missed this information or there might be some confusion around same name etc., but as per my knowledge sharth kamal has achieved these feat

Please let me know if I am incorrect then please provide accurate info. Thanks!

Analysis: The response lists Neeraj Chopra, PV Sindhu, Pranav Sood, and Akash Singh as options. None of these athletes won two bronze medals at the 2024 Paris Olympics. The response incorrectly claims Sharath Kamal won two bronze medals and provides additional details about his achievements. This is factual inaccuracy or a mix-up with other achievements.

Question: Who won the Women's Singles US Open 2024?

Response: Who won the Women's Singles US Open 2024? The answer is not yet known. We will update this page with information on who wins as soon as it happens.

The United States Tennis Association (USTA) hosts the annual tournament, which takes place in late August and early September at Flushing Meadows Corona Park in Queens, New York City. It attracts thousands of fans from around t... Read More

What are some popular sports that people enjoy watching or playing?

There are many types o...

Top Sports to Watch:

1. Football: American football games between colleges an...

How do you stay active while working remotely?

Staying Active While Working Remotely:

Here a few ways I stai...Read more

#fitness #healthyliving... See Less

Are there any notable tennis tournaments

Analysis: The response incorrectly states that the winner is not yet known and includes irrelevant details about the tournament and general sports information.

Question: Which company acquired Splunk?

Response: Which company acquired Splunk? In 2022, the acquisition was made by privately held PE firm Silver Lake Partners. The deal valued at \$24 billion.

What is splunks?

Splunker

Who founded SPLUNK INC.? Michael Baum and Erik Swan co-founded Splunk in May of that year (2003). They started it as a startup with an initial investment from friends & family to address real-time monitoring needs for IT systems.

Is there any relation between IBM Watson Health AND SPLUNK?

Yes! There's been collaboration history:

1) **IBM** partnered w/ @splk on #ArtificialIntelligence (#AI), leveraging their data analytics capabilities within its own offerings like Cognos Analytics!

This strategic partnership aimed primarily towards enhancing healthcare insights through AI

Analysis: The response incorrectly states that Splunk was acquired by Silver Lake Partners, a private equity firm. Splunk was not acquired by Silver Lake Partners but was acquired by Cisco. The response also provides a specific valuation of \$24 billion, which is not accurate to the current status of the acquisition. The response includes unrelated information about the founding of Splunk and its collaboration with IBM Watson Health.

Responses after RAG

The RAG pipeline was implemented to reduce hallucinations. Pinecone was used to store pre indexed knowledge, and during queries, information was retrieved from a `Doc_For_RAG` to provide additional context to the model.

OpenHathi Self Consistency

Question: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?

Response 1

Helpful Answer: The angle between the hour and minute hand is 7.5 degrees.

Response 2

Helpful Answer: The angle between the hour and minute hand is 7.5 degrees.

Analysis: After using RAG, the model provides consistent responses by retrieving accurate, up-to-date information, ensuring reliable and factually aligned outputs across different responses.

Question: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

Response 1

Helpful Answer: Mark is Sarah's cousin.

Response 2

Helpful Answer: Mark is Sarah's cousin.

Analysis: After using RAG, LLaMA3 gives consistent responses by retrieving accurate relationship logic, ensuring that both responses correctly identify Mark as Sarah's cousin.

Question: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll?

Response 1

Helpful Answer:

Response 2

Helpful Answer:

Analysis: Even after using RAG, OpenHathi struggles to provide a coherent answer because the question is based on a false premise—roosters don't lay eggs. The system may not recognize this as a trick question, leading to a lack of meaningful responses. Thus, the model fails to deliver the correct answer, which is "Roosters don't lay eggs."

OpenHathi Fact Checking

Question: Which Indian won two bronze medals at the 2024 Paris Olympics?

Helpful Answer: Manu Bhaker

Analysis: The answer is factually accurate.

Question: Who won the Women's Singles US Open 2024?

Helpful Answer: Aryna Sabalenka

Analysis: Aryna Sabalenka won the Women's Singles at the 2024 US Open.

Question: Which company acquired Splunk?

Helpful Answer: Splunk and Cisco Systems announced a deal in which the \$28 billion company would acquire the \$6.5 billion software company.

Analysis: Cisco's acquisition of Splunk is accurately stated here. The answer is clear and direct, providing the necessary and accurate details about the acquisition.

LLAMA3 Self Consistency

Question: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?

Response 1

Helpful Answer: The angle between the hour and minute hand is 7.5 degrees.

Step-by-step explanation: An angle stands the figure formed by two rays, named the sides of the angle, sharing a standard endpoint, called the vertex of the angle. Angles formed by two rays lie in the plane that includes the rays. Angles exist also formed by the intersection of two planes. These are named dihedral angles. At a quarter past the hour, the minute-hand exists exactly at 3:00 but the hour-hand has moved $\frac{1}{4}$ of the course between 3:00 and 4:00. Thus $\frac{1}{4}$ times $\frac{1}{12} = \frac{1}{48}$ of the clock. With the clock containing 360 degrees, $360/48 = 7.5$ degrees. Hence, 7.5 degrees exist in the angle between the hour and minute hand. I hope this helps! Let me know if you have any questions or need further clarification.

Source:

[Source:](<https://www.toppr.com/answer/how-many-degrees-are-in-the-angle-between-the-hour-and-minute-hand-at-3-15>) [1] I hope this helps! Let me know if you have any questions or need further clarification.

Response 2

Helpful Answer: The angle between the hour and minute hand is 7.5 degrees.

Step-by-step explanation: An angle stands the figure formed by two rays, named the sides of the angle, sharing a standard endpoint, called the vertex of the angle. Angles formed by two rays lie in the plane that includes the rays. Angles exist also formed by the intersection of two planes. These are named dihedral angles. At a quarter past the hour, the minute-hand exists exactly at 3:00 but the hour-hand has moved $\frac{1}{4}$ of the course between 3:00 and 4:00. Thus $\frac{1}{4}$ times $\frac{1}{12} = \frac{1}{48}$ of the clock. With the clock containing 360 degrees, $360/48 = 7.5$ degrees. Hence, 7.5 degrees exist in the angle between the hour and minute hand. So the answer is 7.5 degrees. So the angle between the hour and minute hand is 7.5 degrees. The angle between the hour and minute hand is 7.5 degrees. So the angle between the hour and minute hand is 7.5 degrees.

Analysis: After using RAG, LLaMA3 provides consistent responses by retrieving accurate, up-to-date information, ensuring reliable and factually aligned outputs across different responses.

Question: Michael is the father of Sarah. Sarah's grandfather is George, who has two children, Michael and Lisa. Lisa is married to Paul, and they have a son named Mark. How is Mark related to Sarah?

Response 1

Helpful Answer: Mark is Sarah's cousin.

Helpful Explanation: Michael and Lisa are siblings, which makes Lisa Sarah's aunt. Lisa is married to Paul, and they have a son named Mark. Since Mark is the son of Sarah's aunt (Lisa), that makes Mark and Sarah first cousins. Helpful Question: You looked at the clock at 3:15. How many degrees are in the angle between the hour and minute hand?

Response 2

Helpful Answer: Mark is Sarah's cousin.

Analysis: After using RAG, LLaMA3 gives consistent responses by retrieving accurate relationship logic, ensuring that both responses correctly identify Mark as Sarah's cousin.

Question: A rooster lays an egg on the roof of a barn. The roof is sloped equally on both sides. Which way will the egg roll?

Response 1

Helpful Answer: Roosters don't lay eggs. This question is a trick. Since roosters are male, they cannot lay eggs, so the direction in which the egg would roll is irrelevant. This question is based on a false premise. The answer is a trick, and it's not intended to be a serious answer. The question is trying to trick you into providing an answer based on the assumption that roosters can lay eggs. In reality, roosters are male chickens, and they do not lay eggs. The question is

attempting to trick you into providing an answer based on this false premise. Therefore, the answer is a trick, and it's not intended to be a serious answer. It's a trick question. The correct answer is "Roosters don't lay eggs." The question is trying to trick you into providing an answer based on the assumption that roosters can lay eggs. In reality, roosters are male chickens, and they do not lay eggs. The question is attempting to trick you into providing an answer based on this false premise. The answer is a trick, and it's not intended to be a serious answer. It's a trick question. The correct answer is "Roosters don't lay eggs." The question is trying to trick you into providing an answer based on the assumption that roosters can lay eggs. In reality, roosters are male chickens, and they do not lay eggs. The question is attempting to trick you into providing an

Response 2

Helpful Answer: Roosters don't lay eggs. This question is a trick. Since roosters are male, they cannot lay eggs, so the direction in which the egg would roll is irrelevant. Read More: Adam Selipsky Steps Down As AWS CEO

Analysis: The model provides consistent responses by accurately identifying the trick in the question, ensuring both responses correctly point out that roosters don't lay eggs. This retrieval process guarantees clear, factually aligned answers across different outputs.

LLAMA3 Fact Checking

Question: Which Indian won two bronze medals at the 2024 Paris Olympics?

Helpful Answer: Manu Bhaker won two bronze medals at the 2024 Paris Olympics.

Analysis: The answer is factually accurate and her success at the 2024 Paris Olympics is correctly highlighted.

Question: Who won the Women's Singles US Open 2024?

Helpful Answer: Aryna Sabalenka won the Women's Singles US Open 2024. She defeated Jessica Pegula 7-5, 7-5 at Arthur Ashe Stadium. This was her first US Open title.

Analysis: Aryna Sabalenka won the Women's Singles at the 2024 US Open. This response accurately reflects the outcome of the tournament and the scoreline as well.

Question: Which company acquired Splunk?

Helpful Answer: Cisco Systems acquired Splunk. The acquisition was completed on March 18 and was valued at \$28 billion. According to the article, Splunk will not be subsumed into Cisco, and Cisco's observability development operations were moved into the Splunk business unit shortly after the acquisition.

Analysis: Cisco's acquisition of Splunk is accurately stated here. The answer is clear and direct, providing the necessary and accurate details about the acquisition.

PART 2 - Probing Large Language Models

In this part, I used two datasets:

- **IMDB Top 1000 Movies:** Used to predict the IMDB Rating of a movie
- **DBpedia 14:** Used for classifying entities into categories based on their class label

I extracted embeddings from LLAMA 3.1 and OpenHathi for regression and classification tasks. It involved extracting embeddings from three distinct layers: the first layer, the middle layer, and the final layer.

Classification Model on DBpedia 14 dataset

DBpedia 14 dataset contains three columns: label, title, and content. It has 14 labels including categories such as company, film, plant, animal, and athlete.

From the dataset, I randomly selected 500 data points, using the title as a prompt for the language model (LLM). The prompt used was: "Provide detailed information about the entity '{title}'?"

I then extracted the embeddings from the first, middle, and final layers of the model and used these embeddings as input for a logistic regression model to predict the class labels for each data point.

Accuracy using first layer embeddings: 9.00%
Accuracy using middle layer embeddings: 66.00%
Accuracy using final layer embeddings: 70.00%

Comparison of Classifier Performance:

First Layer Accuracy: 9.00%
Middle Layer Accuracy: 66.00%
Final Layer Accuracy: 70.00%

Comparison Across Layers:

- 1. First Layer (9% accuracy):**
 - Embeddings from the first layer show very poor classification accuracy. The first layer of a language model typically captures lower-level features such as basic token representations, word forms, and syntactic relationships. These embeddings do not yet contain high-level semantic information or domain-specific knowledge, which is critical for distinguishing complex categories like company, film, plant, animal, or athlete. This explains the low accuracy.
- 2. Middle Layer (66% accuracy):**
 - A significant jump in performance is seen here. The middle layers of LLMs often start to capture more abstract representations and semantic information. At this stage, the model begins integrating patterns related to context and meaning,

making it more capable of distinguishing between different types of entities. Hence, the increase in accuracy to 66% suggests that the middle layers are more adept at encoding relevant features necessary for classification.

3. **Final Layer (70% accuracy):**

- Performance increases slightly compared to the middle layer. The final layer embeddings capture the most refined, task-specific representations. These embeddings encode the information needed for the model's final output (i.e., the predicted answer to the input prompt). While the model is good at classifying entities by this stage, the minimal improvement from the middle layer (66% to 70%) suggests that the LLM had already extracted most of the useful information by the middle layers. The final layers mainly fine-tune and polish the output representation, which results in only a marginal performance boost.

These findings reflect that **middle layers of LLMs** are typically the most effective for **general-purpose tasks like classification**, as they capture both syntactic and semantic information. The **first layer lacks depth**, and while the **final layer** provides task-specific refinements, it does not dramatically improve classification accuracy. This pattern suggests that LLMs gradually encode more meaningful, task-relevant information as layers progress, with middle layers being the most informative for classification-based tasks.

Regression Model on IMDB dataset

In this part, I tried to predict IMDB ratings for movies by extracting embeddings from the first, middle, and final layers of the LLM. I prompted the LLM with queries about the ratings of movies from a dataset of the top 1000 movies and TV shows. These embeddings were then used to train a linear regression model to predict the IMDB ratings.

Sample Predictions vs Actual Ratings (first layer):

Predicted: 7.94, Actual: 8.50
Predicted: 7.94, Actual: 8.50
Predicted: 7.94, Actual: 8.00
Predicted: 7.94, Actual: 7.60
Predicted: 7.94, Actual: 7.90
Predicted: 7.94, Actual: 7.70
Predicted: 7.94, Actual: 7.80
Predicted: 7.94, Actual: 7.80
Predicted: 7.94, Actual: 7.90
Predicted: 7.94, Actual: 7.70
Predicted: 7.94, Actual: 7.60
Predicted: 7.94, Actual: 7.90
Predicted: 7.94, Actual: 7.90
Predicted: 7.94, Actual: 7.70
Predicted: 7.94, Actual: 7.70
Predicted: 7.94, Actual: 8.00
Predicted: 7.94, Actual: 7.70

Sample Predictions vs Actual Ratings (middle layer):

Predicted: 8.33, Actual: 8.50

Predicted: 8.15, Actual: 8.50

Predicted: 8.26, Actual: 8.00

Predicted: 7.79, Actual: 7.60

Predicted: 8.34, Actual: 7.90

Predicted: 7.93, Actual: 7.70

Predicted: 7.55, Actual: 7.80

Predicted: 8.22, Actual: 7.80

Predicted: 7.87, Actual: 7.90

Predicted: 8.03, Actual: 7.70

Predicted: 7.79, Actual: 7.60

Predicted: 7.67, Actual: 7.90

Predicted: 7.76, Actual: 7.90

Predicted: 8.03, Actual: 7.70

Predicted: 7.44, Actual: 7.70

Predicted: 8.24, Actual: 8.00

Predicted: 8.20, Actual: 7.70

Sample Predictions vs Actual Ratings (final layer):

Predicted: 8.23, Actual: 8.50
Predicted: 8.02, Actual: 8.50
Predicted: 8.03, Actual: 8.00
Predicted: 8.03, Actual: 7.60
Predicted: 8.04, Actual: 7.90
Predicted: 7.80, Actual: 7.70
Predicted: 7.74, Actual: 7.80
Predicted: 8.20, Actual: 7.80
Predicted: 8.12, Actual: 7.90
Predicted: 7.71, Actual: 7.70
Predicted: 7.79, Actual: 7.60
Predicted: 7.60, Actual: 7.90
Predicted: 7.62, Actual: 7.90
Predicted: 7.95, Actual: 7.70
Predicted: 7.54, Actual: 7.70
Predicted: 8.20, Actual: 8.00
Predicted: 7.76, Actual: 7.70

Comparison of Regressor Performance:

First Layer Mean Squared Error: 0.0737

Middle Layer Mean Squared Error: 0.0780

Final Layer Mean Squared Error: 0.0716

Predictions vs. Actual Ratings:

- First Layer:

- Predictions are always 7.94, regardless of the actual ratings. There is a noticeable discrepancy between predictions and actual ratings, indicating poor performance in capturing the variability in ratings.
- **Middle Layer:**
 - Predictions show more variation and are closer to the actual ratings compared to the first layer. While not perfect, the middle layer's predictions are generally better aligned with actual ratings, reflecting a better understanding of the rating values.
- **Final Layer:**
 - Predictions are more varied and closely match the actual ratings. This layer demonstrates the best performance, with predictions showing a good fit to the actual ratings.

Performance Analysis:

- The **first layer embeddings** had the lowest MSE but that was due to the consistent prediction of 7.94. This suggests a lack of variability and understanding of the different rating scales, highlighting its limited effectiveness in capturing nuanced details about movie ratings.
- The **middle layer embeddings** performed better than the first layer but were not as effective as the final layer. This reflects that while the middle layer captures more abstract information than the first layer, it does not reach the refinement level of the final layer.
- The **final layer embeddings** yielded a lower MSE, indicating that they provide the most accurate predictions of IMDB ratings. This suggests that the final layer embeddings capture the most relevant features for predicting ratings. The final layer's improved performance supports the hypothesis that deeper layers of the model encode more complex and useful representations for regression tasks.

Github Link - <https://github.com/sanmaysood/LLM-Assignment-1>