# Q1 : EXPLORING

Fact Checking: (Llama)

1.)"who was the first black scientist in Brazil"

2.)”Name a one eyed animal?”

3.)”Which animal walks on water?”

(Openhaathi)

1.)Which animal goes with 3 legs up the mountain and comes down with 4 legs

2.)“वन प्लस नाइन्टीसिक्स पोंट panch क्या है?”

3.)Which animal has one leg

Self-Consistency (Llama)

1.)"Multiply the smallest positive integer smaller than 100 by the largest negative integer greater than -100"

2.)”Count the number of occurrences of the letter ’L’ in the word -’LOLLAPALOOZA’.”

3.)”spell the reverse of lollipop”

Report

Factual hallucinations: The model occasionally makes up facts, especially when asked highly specific questions that require external knowledge.

Consistency hallucinations: Sometimes, the model misunderstands the context, resulting in answers that don’t align with the user’s prompt. This often happens when the prompt is vague or unclear, causing the model to fill in gaps with incorrect assumptions.

Note: OpenHathi tends to produce non-standard or incorrect Hindi phrases, particularly with unclear prompts. It can generate grammatically incorrect or overly literal translations from English to Hindi, leading to language-specific errors.

Llama also occasionally hallucinates while solving simple math problems!

# Q2 : PROBING

A) Dataset Prepration

I have chosen **2 datasets - one for regression and one for classification**

Regression- Housing price dataset - kaggle

- Converted Features into a prompt to predict price:
- Example prompt:- "This House has an area of 7420 sqft. It has 4 bedrooms and 2 bathrooms . It is a 3 storey house on the mainroad. The house is furnished and

comes with facilities like , air conditioning along with 2 parking slots. It is located in a preferred area. What is the price of this house?"

- Crafted prompts - regression_prep.ipynb => newregression.csv

Classification - Text classification - kaggle

- Put Text as it is to form prompts. Use prompts to further classify text into 5 classes
- 5 classes : Politics = 0 Sport = 1 Technology = 2 Entertainment =3 Business = 4

B) Evaluation and Discussion:

**Regression:** The final Layer is very well able to predict prices . After scaling ,model doesn't seem to overfit .

First layer MSE: 2.1109
Middle layer MSE :1.8991
Final Layer MSE: 1.1552

The embeddings from the final layer yield the best performance, suggesting that the LLM enhances the data representation as it advances through its layers. This progressive refinement likely captures more relevant semantic and contextual details, aiding in more accurate predictions.

The first layer, with the highest MSE, indicates that early representations are less informative or refined. These embeddings are probably closer to the raw input, missing the deeper abstraction needed for accurate regression tasks.

Performance from the middle layer shows improvement compared to the first layer but still falls short of the final layer. This is expected, as the model continues to refine its data understanding through its successive layers.

The observed pattern illustrates the Layer Hierarchy in LLMs. These results demonstrate how LLMs encode information, where early layers capture more superficial details (like syntax) while deeper layers acquire more abstract and task-specific features. This aligns with the established behavior of transformer-based models.

**Classification:**
First layer Accuracy: 42.13%
Middle layer Accuracy :64.43%
Final Layer Accuracy: 89.93%

The classifier's performance notably improves from the first layer to the final layer, reflecting the typical behavior of transformer models like GPT and BERT. Early layers tend to focus on fundamental, general features, while later layers are more adept at capturing complex, task-specific details.

For instance, the accuracy of the first layer is relatively low at 42.13%, suggesting that it mainly processes very basic features, such as word-level or syntactic patterns, which may not be highly effective for tasks requiring deeper semantic understanding.

In contrast, the accuracy for the middle layer rises to 64.43%, indicating that by this stage, the model begins to identify more meaningful patterns and representations that are better suited for classification tasks.

By the final layer, accuracy reaches 89.93%, demonstrating that the embeddings at this stage contain rich, task-relevant information. At this point, the model has likely integrated both syntactic structure and semantic meaning, leading to strong classification predictions.

These results support the notion that deeper layers of the LLM become more specialized for tasks involving contextual understanding and semantic features. The final layer embeddings, being the most refined, excel in classification tasks.

Pattern found describes Layer Hierarchy in LLMs.These findings reflect how LLMs encode information. Early layers capture more surface-level information (like syntax), while deeper layers learn more abstract and task-specific features. This is a well-known characteristic of transformer-based models.