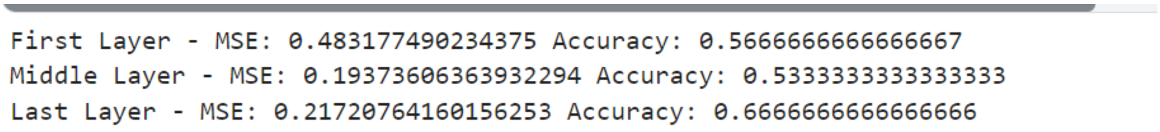
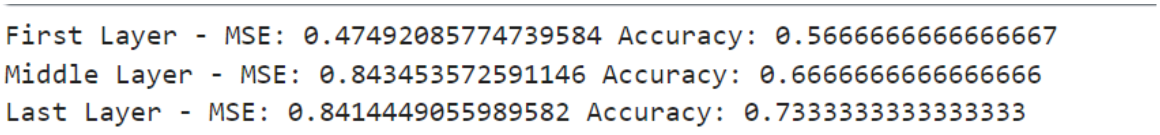
OpenLlama



```
First Layer - MSE: 0.483177490234375 Accuracy: 0.5666666666666667
Middle Layer - MSE: 0.19373606363932294 Accuracy: 0.5333333333333333
Last Layer - MSE: 0.21720764160156253 Accuracy: 0.6666666666666666
```

OpenHathi



```
First Layer - MSE: 0.47492085774739584 Accuracy: 0.5666666666666667
Middle Layer - MSE: 0.843453572591146 Accuracy: 0.6666666666666666
Last Layer - MSE: 0.8414449055989582 Accuracy: 0.7333333333333333
```

Discussion

I have used the action.csv file of action movies and their ratings in **imdb-movies-dataset-based-on-genre dataset.** I have regressed to predict the movie ratings and used the certification of the film as the categorical variable to predict.

Classification

Both models show significant increase in the accuracy from first layer to the last layer.
This means the last layer captures better information and knowledge for classification tasks. These could be some reasons for this:

- In deep learning models, particularly in transformer-based models, early layers typically capture lower-level features (like general syntax or local structures), while later layers refine these features into more abstract, high-level representations.
- According to this principle, the deeper layers in a network tend to compress the input information while retaining only the most relevant parts for the target task. Since we prompt to query the rating and certification for the model, it is likely that the last layer captures this information.

- The first layer's representation may still be relatively raw, and the middle layer may reflect intermediate processing. The last layer benefits from all previous transformations, which could be critical for complex tasks like classification.

Regression

We observe the MSE growing from first layer to final layer in case of OpenHathi which is both undesirable and unexpected. This is an anomaly and indicates the last layer does not capture appropriate information for the regression task.

LLMs are trained to capture the semantic meaning and relationships between words and phrases rather than precise numeric outputs. As you move toward deeper layers, the model focuses more on understanding the broader context and meaning of the text. In this process, the representations might lose the precise numeric information that is necessary for tasks like regression.
The last layers of LLMs often compress or abstract the information from earlier layers to make predictions. This compression process is more suited to tasks where categorical decision-making is required (such as classification) and less useful for tasks requiring precise numerical outputs. The model may collapse certain details important for regression, hence increasing MSE.

In case of LlaMa there is a decrease in MSE from first to middle layer and then a slight increase. This shows that LLaMa is able to capture the knowledge required for predicting the rating and does not compress this information in the earlier layers but rather amplifies this.

Across Models

The anomaly is the slightly better performance of the Hathi model on the prediction of the movie certification than LLaMa on this task, this is unexpected as the dataset contains mostly Hollywood movies while OpenHathi has been trained primarily on Indian data.
However OpenLLaMa performs better on predicting the rating of the movie.