## Discussion for questions 2 and 3:

**If you recall the paper we reviews on consistency checking used several models, do you think we can use consistency check method between these layers for factuality analysis? Present your approach and results including discussion.**

In the consistency checking paper, it seems that LLaMA2 adjusts its predictions in the higher layers when factual knowledge is needed, the change in probabilities is instantaneous, from an almost zero probability to one which is very high. During human testing, I observed that token selection sometimes became clear around layer 23, with significantly higher predictions compared to lower layers, where probabilities were close to zero. This poses a challenge, as it can occasionally lead to incorrect token selection towards the end of the process, favoring tokens with steeply higher changes over potentially more factual tokens with slightly lower probabilities at the end layer.

Contrasting the layers before and after a change can boost the knowledge emerging from the higher layers and encourage the model to rely more on factual information. To prioritize the mature layer, we can utilize the contrastive method from the paper. This involves comparing predictions from both layers and subtracting less reliable predictions from more reliable ones. Then, we use this comparison to determine the next word in a sentence. This approach splits from conventional methods where probability predictions are only based on the final layer's output. Contrasting the output distributions of a selected "premature" layer with those of the final layer can facilitate a better understanding of token probabilities.

**Write another discussion explaining the how the layers effect on the different metrics on your trained model from assignment 1.c.**

In my testing, I observed a consistent trend: as the layers increased, so did the metrics. However, it wasn't until layer 24 that I noticed a significant change. At the final layer of the Llama 2 model, I achieved notable improvements in BLEU, Rouge-L, BERTScore, and CodeBLEU.

Furthermore, during human testing, which is more convenient, I noted that the correct token selection sometimes became apparent around layer 23. However, as I progressed through subsequent layers, the probability of selecting the correct token fluctuated. Sometimes it decreased, other times it increased, and occasionally it remained unchanged. This inconsistency poses a challenge as it occasionally leads to incorrect token selection towards the end of the process.