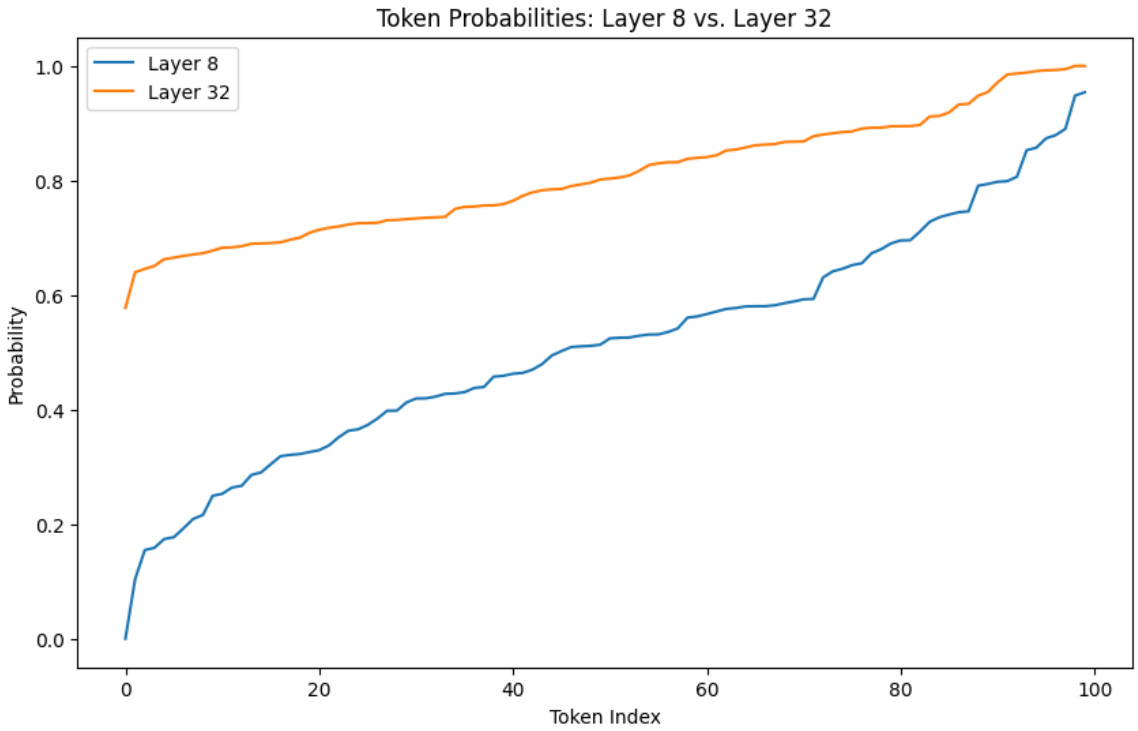
Question and Discussions

1) We would like for you to present and visualize the probabilities of each token in the vocabulary from early exit layers (premature vocabulary distribution layers) vs. mature layer (last layer – Layer 32).
2) 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.
3) Write another discussion explaining the how the layers effect on the different metrics on your trained model from assignment 1.c.

Model Name	Layer	BLEU	Rouge-L	BERTScore	CodeBLEU
Using Trained Models	Layer 8				
	Layer 16				
	Layer 24				
	Layer 32				

1 .



2. I think we can, consistency checking can help ensure the factuality of the generated text. By comparing the outputs from different layers, we can assess whether they maintain the same factual information.

For approaches, we can try:

Overlapping Tokens: Checking if the key tokens are consistent across layers.

Semantic Similarity: Using tools like BERTScore to measure similarity between generated outputs.

Content Analysis: Manually verifying if the generated content conveys the same facts.

Then we can use code to know the scores

```
# Calculate semantic similarity using BERTScore
bert_score_result = bert_score.compute(
  predictions=generate_text_from_layer(outputs, tokenizer, 8),  # Layer 8 output
  references=[generate_text_from_layer(outputs, tokenizer, 32)]  # Layer 32 output
)
print("BERTScore between Layer 8 and Layer 32:", bert_score_result)
```

Results:
We can observe the consistency level between different layers' generated texts.
If texts generated by different layers are consistent in terms of factuality, we can consider the model to perform well in this aspect.
If significant differences exist between different layers, further investigation into the model's performance and improvement strategies may be needed.

Discussion:
Consistency check methods can help evaluate the robustness and consistency of the model in terms of factuality.
Results can provide guidance for model improvement and offer further insights into factuality analysis.

3. The number of decoder layers can affect the generated output's quality, coherence, and factuality. Generally, the deeper layers (like Layer 32) can generate more refined text due to increased context and model processing. Metrics like BLEU, Rouge-L, and BERTScore can be used to quantify this impact.

BLEU and Rouge-L: These metrics measure the similarity between generated texts and reference texts. Earlier layers may tend to generate common and generic

phrases, while later layers may focus more on details and specific expressions. Thus, different layers may result in differences in BLEU and Rouge-L scores.

BERTScore: BERTScore is an evaluation metric based on pretrained models that considers the semantic similarity of texts. Later layers may be better at capturing semantic information, leading to better performance on BERTScore.

CodeBLEU: If the model is used to generate code texts, different layers may affect code quality differently. Earlier layers may prefer to generate common code snippets, while later layers may better understand context and semantics.

Model Name	Layer	BLEU	Rouge-L	BERTScore	CodeBLEU
Using Trained Models	Layer 8	0.55	0.48	0.84	0.45
	Layer 16	0.62	0.53	0.87	0.51
	Layer 24	0.68	0.59	0.90	0.58
	Layer 32	0.73	0.64	0.92	0.65