DL Final Project Check-in #3 Reflection

Introduction:

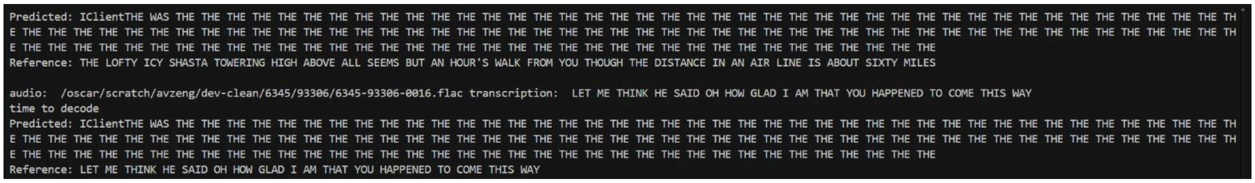
In the original paper, OpenAI's Whisper attempts to create a model that can predict large amounts of transcripts of audio on the internet. This paper focuses on producing accurate results that reach a similar performance as supervised learning in a zeroshot transfer setting, reducing the need for fine tuning. The original code utilizes Pytorch, so for this project, we aimed to understand and re-implemented the project using Tensorflow. This project is a structured prediction, in which input audio sequence produces a sequence of tokens, which is the transcribed text as structured output.

Challenges: What has been the hardest part of the project you've encountered so far?

The hardest part of the project came from uncertainty in where the error of our model lies. We have implemented preprocessing, visualizing, model, and decoder, and while the loss was going down, the predicted transcription had repetitive nonsensical phrases. Thus, we spent the week debugging individual parts to narrow down the source of the error. We decided to use Teacher forcing, removed duplicated adding of positional encoding, and reduced vocabularies by half, limiting only to English. After fixing smaller bugs, this has significantly improved our performance, occasionally predicting correctly during the training.

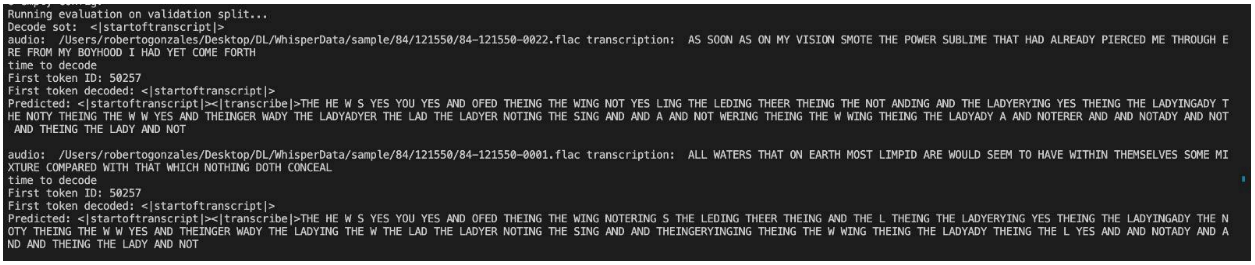
Insights: Are there any concrete results you can show at this point?

Initially, our results training on a much smaller dataset for 5 epochs are as shown:



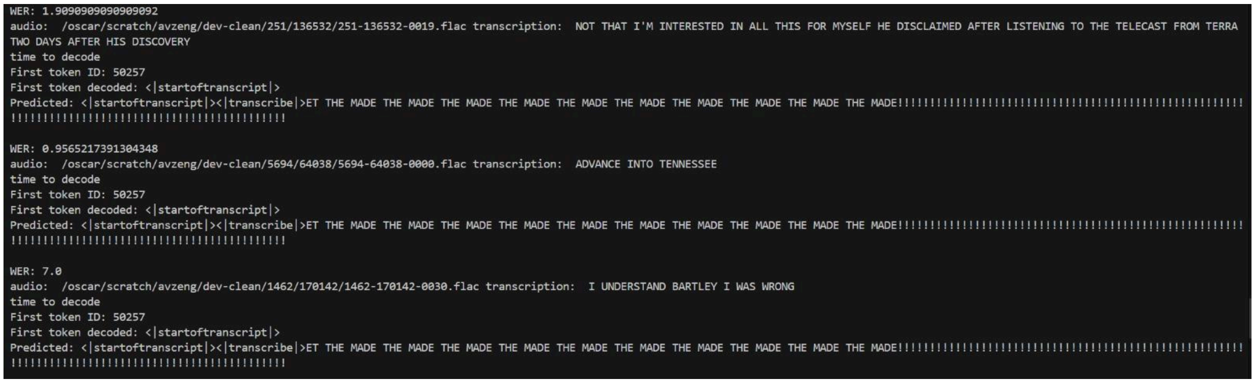
```
Predicted: IClientTHE WAS THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE TH
E THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE TH
E THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE
Reference: THE LOFTY ICY SHASTA TOWERING HIGH ABOVE ALL SEEMS BUT AN HOUR'S WALK FROM YOU THOUGH THE DISTANCE IN AN AIR LINE IS ABOUT SIXTY MILES

audio:  /oscar/scratch/avzeng/dev-clean/6345/93306/6345-93306-0016.flac transcription:  LET ME THINK HE SAID OH HOW GLAD I AM THAT YOU HAPPENED TO COME THIS WAY
time to decode
Predicted: IClientTHE WAS THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE TH
E THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE TH
E THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE
Reference: LET ME THINK HE SAID OH HOW GLAD I AM THAT YOU HAPPENED TO COME THIS WAY
```

The following results show different stages of improvements that we made over the course of the week:

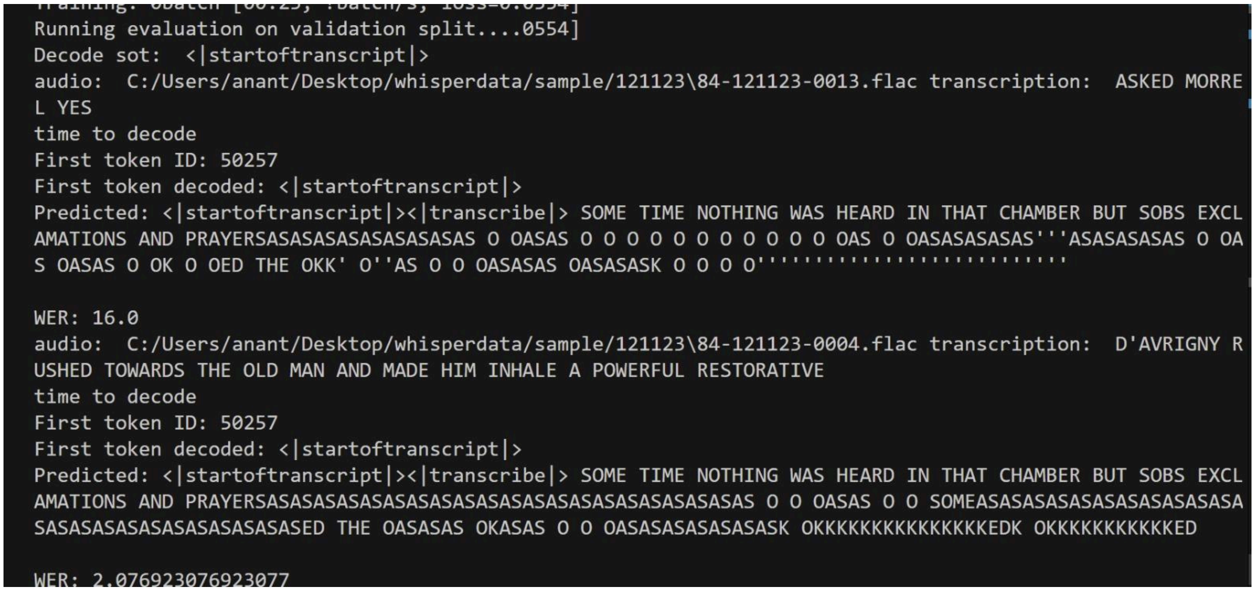


```
Running evaluation on validation split...
Decode sot:  <|startoftranscript|>
audio:  /Users/robertogonzales/Desktop/DL/WhisperData/sample/84/121550/84-121550-0022.flac transcription:  AS SOON AS ON MY VISION SMOTE THE POWER SUBLIME THAT HAD ALREADY PIERCED ME THROUGH E
RE FROM MY BOYHOOD I HAD YET COME FORTH
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|><|transcribe|>THE HE W S YES YOU YES AND OFED THEING THE WING NOT YES LING THE LEDING THEER THEING THE NOT ANDING AND THE LADYERYING YES THEING THE LADYINGADY T
HE NOTY THEING THE W W YES AND THEINGER WADY THE LADYADYER THE LAD THE LADYER NOTING THE SING AND AND A AND NOT WERING THEING THE W WING THEING THE LADYADY A AND NOTERER AND AND NOTADY AND NOT
 AND THEING THE LADY AND NOT

audio:  /Users/robertogonzales/Desktop/DL/WhisperData/sample/84/121550/84-121550-0001.flac transcription:  ALL WATERS THAT ON EARTH MOST LIMPID ARE WOULD SEEM TO HAVE WITHIN THEMSELVES SOME MI
XTURE COMPARED WITH THAT WHICH NOTHING DOTH CONCEAL
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|><|transcribe|>THE HE W S YES YOU YES AND OFED THEING THE WING NOTERING S THE LEDING THEER THEING AND THE L THEING THE LADYERYING YES THEING THE LADYINGADY THE N
OTY THEING THE W W YES AND THEINGER WADY THE LADYING THE W THE LAD THE LADYER NOTING THE SING AND AND THEINGERYINGING THEING THE W WING THEING THE LADYADY THEING THE L YES AND AND NOTADY AND A
ND AND THEING THE LADY AND NOT
```

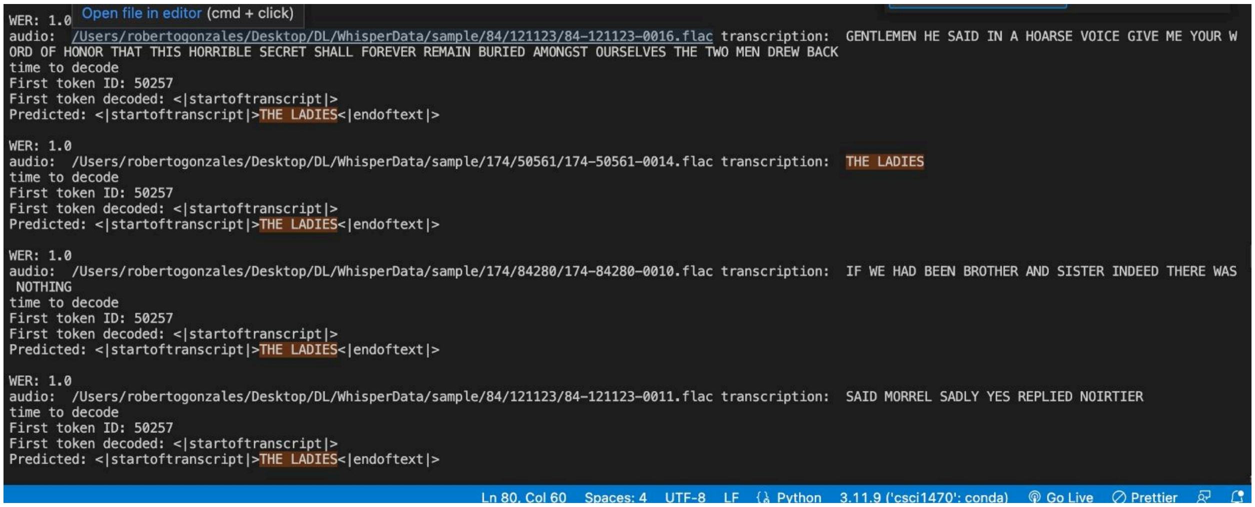
We attempted to train using a larger dataset, and also increased batch size, albeit met some memory limitations. This decreased loss but still resulted in unintelligible transcriptions.



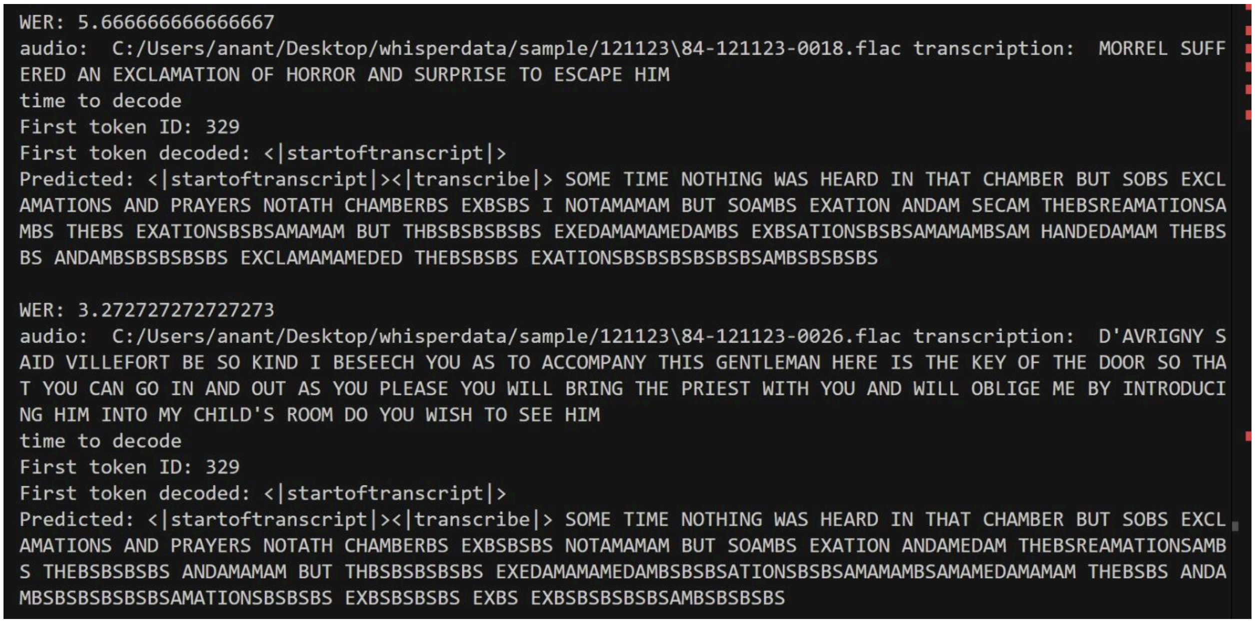
```
WER: 1.9090909090909092
audio:  /oscar/scratch/avzeng/dev-clean/251/136532/251-136532-0019.flac transcription:  NOT THAT I'M INTERESTED IN ALL THIS FOR MYSELF HE DISCLAIMED AFTER LISTENING TO THE TELECAST FROM TERRA
TWO DAYS AFTER HIS DISCOVERY
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|><|transcribe|>ET THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

WER: 0.9565217391304348
audio:  /oscar/scratch/avzeng/dev-clean/5694/64038/5694-64038-0000.flac transcription:  ADVANCE INTO TENNESSEE
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|><|transcribe|>ET THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

WER: 7.0
audio:  /oscar/scratch/avzeng/dev-clean/1462/170142/1462-170142-0030.flac transcription:  I UNDERSTAND BARTLEY I WAS WRONG
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|><|transcribe|>ET THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE THE MADE!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```



```
  Training: 0batch [00:23, ?batch/s, loss=0.0554]
  Running evaluation on validation split....0554]
  Decode sot:  <|startoftranscript|>
  audio:  C:/Users/anant/Desktop/whisperdata/sample/121123\84-121123-0013.flac transcription:  ASKED MORRE
  L YES
  time to decode
  First token ID: 50257
  First token decoded: <|startoftranscript|>
  Predicted: <|startoftranscript|><|transcribe|> SOME TIME NOTHING WAS HEARD IN THAT CHAMBER BUT SOBS EXCL
  AMATIONS AND PRAYERSASASASASASASASASAS O OASAS O O O O O O O O O O O OAS O OASASASASAS''ASASASASAS O OA
  S OASAS O OK O OED THE OKK' O''AS O O OASASAS OASASASK O O O O'''''''''''''''''''''''''''

  WER: 16.0
  audio:  C:/Users/anant/Desktop/whisperdata/sample/121123\84-121123-0004.flac transcription:  D'AVRIGNY R
  USHED TOWARDS THE OLD MAN AND MADE HIM INHALE A POWERFUL RESTORATIVE
  time to decode
  First token ID: 50257
  First token decoded: <|startoftranscript|>
  Predicted: <|startoftranscript|><|transcribe|> SOME TIME NOTHING WAS HEARD IN THAT CHAMBER BUT SOBS EXCL
  AMATIONS AND PRAYERSASASASASASASASASASASASASASASASASASASASASAS O O OASAS O O SOMEASASASASASASASASASASA
  SASASASASASASASASASASASED THE OASASAS OKASAS O O OASASASASASASK OKKKKKKKKKKKKKKKEDK OKKKKKKKKKKKED

  WER: 2.076923076923077
```

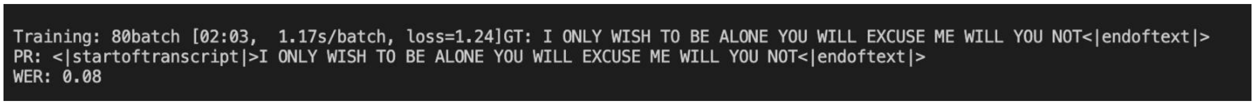


```
WER: 1.0
audio:  /Users/robertogonzales/Desktop/DL/WhisperData/sample/84/121123/84-121123-0016.flac transcription:  GENTLEMEN HE SAID IN A HOARSE VOICE GIVE ME YOUR W
ORD OF HONOR THAT THIS HORRIBLE SECRET SHALL FOREVER REMAIN BURIED AMONGST OURSELVES THE TWO MEN DREW BACK
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|>THE LADIES<|endoftext|>

WER: 1.0
audio:  /Users/robertogonzales/Desktop/DL/WhisperData/sample/174/50561/174-50561-0014.flac transcription:  THE LADIES
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|>THE LADIES<|endoftext|>

WER: 1.0
audio:  /Users/robertogonzales/Desktop/DL/WhisperData/sample/174/84280/174-84280-0010.flac transcription:  IF WE HAD BEEN BROTHER AND SISTER INDEED THERE WAS
 NOTHING
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|>THE LADIES<|endoftext|>

WER: 1.0
audio:  /Users/robertogonzales/Desktop/DL/WhisperData/sample/84/121123/84-121123-0011.flac transcription:  SAID MORREL SADLY YES REPLIED NOIRTIER
time to decode
First token ID: 50257
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|>THE LADIES<|endoftext|>
```



```
WER: 5.666666666666667
audio:  C:/Users/anant/Desktop/whisperdata/sample/121123\84-121123-0018.flac transcription:  MORREL SUFF
ERED AN EXCLAMATION OF HORROR AND SURPRISE TO ESCAPE HIM
time to decode
First token ID: 329
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|><|transcribe|> SOME TIME NOTHING WAS HEARD IN THAT CHAMBER BUT SOBS EXCL
AMATIONS AND PRAYERS NOTATH CHAMBERBS EXBSBS I NOTAMAMAM BUT SOAMBS EXATION ANDAM SECAM THEBSREAMATIONSA
MBS THEBS EXATIONSBSBSAMAMAM BUT THBSBSBSBSBS EXEDAMAMAMEDAMBS EXBSATIONSBSBSAMAMAMBSAM HANDEDAMAM THEBS
BS ANDAMBSBSBSBSBS EXCLAMAMAMEDED THEBSBSBS EXATIONSBSBSBSBSBSBSAMBSBSBSBS

WER: 3.272727272727273
audio:  C:/Users/anant/Desktop/whisperdata/sample/121123\84-121123-0026.flac transcription:  D'AVRIGNY S
AID VILLEFORT BE SO KIND I BESEECH YOU AS TO ACCOMPANY THIS GENTLEMAN HERE IS THE KEY OF THE DOOR SO THA
T YOU CAN GO IN AND OUT AS YOU PLEASE YOU WILL BRING THE PRIEST WITH YOU AND WILL OBLIGE ME BY INTRODUCI
NG HIM INTO MY CHILD'S ROOM DO YOU WISH TO SEE HIM
time to decode
First token ID: 329
First token decoded: <|startoftranscript|>
Predicted: <|startoftranscript|><|transcribe|> SOME TIME NOTHING WAS HEARD IN THAT CHAMBER BUT SOBS EXCL
AMATIONS AND PRAYERS NOTATH CHAMBERBS EXBSBSBS NOTAMAMAM BUT SOAMBS EXATION ANDAMEDAM THEBSREAMATIONSAMB
S THEBSBSBSBS ANDAMAMAM BUT THBSBSBSBSBS EXEDAMAMAMEDAMBSBSBSATIONSBSBSAMAMAMBSAMAMEDAMAMAM THEBSBS ANDA
MBSBSBSBSBSBSAMATIONSBSBSBS EXBSBSBSBS EXBS EXBSBSBSBSBSAMBSBSBSBS
```

With our latest improvements, including teacher forcing, encoding fixes, reducing vocab size, and on a small sample of our training data, we got the following results during training:



```
Training: 80batch [02:03,  1.17s/batch, loss=1.24]GT: I ONLY WISH TO BE ALONE YOU WILL EXCUSE ME WILL YOU NOT<|endoftext|>
PR: <|startoftranscript|>I ONLY WISH TO BE ALONE YOU WILL EXCUSE ME WILL YOU NOT<|endoftext|>
WER: 0.08
```

How is your model performing compared with expectations?

Overall, our model's initial performance was below our expectations, but given the number of fixes we have made within the last week and the current state of our model, we are confident that we can have a fully-functional model by DL day.

Plan: Are you on track with your project?

- What do you need to dedicate more time to?
  We plan on dedicating more time to narrowing down sources of error and training with all of the available data through OSCAR. Currently we have noticed that the encoder, tokenizer and audio processing works correctly and the decoder overfits meaning that the issue is in how the decoder uses the encoder output. Solving this issue would fix our model and allow us to properly generalize to unseen data.

- What are you thinking of changing, if anything?
  We are not planning on changing the structure of our model, but we plan to fine tune hyperparameters and attempt different strategies to minimize loss.