For this checkin, we also require you to write up a reflection including the following:

- **Introduction**: This can be copied from the proposal.
- **Challenges**: What has been the hardest part of the project you've encountered so far?
- **Insights**: Are there any concrete results you can show at this point?
  - How is your model performing compared with expectations?
- **Plan**: Are you on track with your project?
  - What do you need to dedicate more time to?
  - What are you thinking of changing, if anything?

This check in meeting with your mentor TA can either be in-person or over Zoom, Google Meet, etc. Reach out to your mentor TA before 04/24 to schedule this meeting.

Regarding what we generally expect you to have **done** by this time:

- You should have collected any data and preprocessed it.
- You should have shared the Github repo link with your mentor TA
- You should have almost finished implementing your model, and are working on training your models and ablation experiments.
- Please make sure you are keeping your list of public implementations you've found up-to-date.

Introduction: We are trying to solve the problem of incomplete genome assemblies. When an organism's genome is sequenced, it must be done in chunks, as current DNA sequencing technologies only work for certain lengths of reads. These reads are then aligned together into a reference genome for that organism. However, there are certain instances where a complete reference genome can not be assembled due to challenges in alignment, such as repetitive regions. As a result, we propose the implementation of a transformer model trained to learn long-range dependencies in DNA sequences, in order to fill these gaps in the genome. The problem we are trying to solve is that of generation: we want our model to learn the underlying distribution of DNA sequence data and be able to generate new sequences.

In larger-scale genomic projects, such as metagenomic studies, sequences can be unlabeled, meaning the species they came from is unknown. Different species have different genomic features, and knowing these unique features can vastly improve the accuracy of a genome assembler. A generalized model that aims to fill in gaps in incomplete assemblies will likely have less accurate results than one that is fine-tuned for that specific species. Thus, we propose adding a classifier prior to the main transformer that generates the missing gaps. The result of this classifier can then be used to fine-tune our main transformer to the specific species the classifier predicted the input sequence to be from.

Challenges: The hardest part about implementing the model was implementing the masking mechanism whereby parts of the input are masked and then the model has to predict the masked tokens. We went through various different implementations, but the way we ended up deciding on was to pass in the full unmasked input to the model along with a vector of booleans with each boolean corresponding to whether the associated position in the input sequence is masked. Now, our biggest challenge is training because we are getting really high accuracy values more than expected, leading us to think that there may be something going wrong with the training and masking process.

Insights: We have generated a k-mer (6-mer) pickled dictionary and preprocessed the first ~7,000,000 bp (100,000 lines) of the GenBank fasta file for chromosome 21 of the T2T-CHM13v2.0 genome assembly, including splitting the fasta file based on a window, tokenization, and k-mer encoding. We have also started training our transformer model using this preprocessed data. Our model does not currently perform according to our expectations because we are reaching very high accuracy. We think that there is an issue with how the model is masking and predicting the gaps, so we will continue troubleshooting to fix our model implementation.

Plan: We are on track. We have finished preprocessing and implementing our model. We need to devote more of our time to training and finding the optimal hyperparameters. We plan to do this by splitting up the parameter space among us four and then testing our assigned parameters. This way, we can accelerate the training/tuning process because each training run currently takes a while to run. We plan to then start designing our poster on Sunday to be finished by DL day. Once we have optimized our model, we plan to scale up our project by using more data, such as longer sequences, other chromosomes, and/or other species (time permitting).