So far in terms of getting ready for analysis, I have looked at testing the model with contrast sets, check lists, and analyzing performance through the “competency problems” framework. Out of these 3, I am interested most in analyzing the robustness of the model against contrast sets.

Using the boilerplate, I currently have a working trained model using the full SNLI dataset to play around with, possibly playing around with a small eval set to see how contrasting corpuses perform on the model currently.

I think I plan on analyzing with one other set besides SNLI (probably MultiNLI) to see if it performs any differently. To get the contrast sets, I will probably first have to see if there are any online resources that can generate them since both datasets are so large.

From there, I will probably start with seeing how fine tuning with these sets can improve the performance before I attempt anything else.