For our final project, we initially decided to work with the Kaggle Spotify dataset (https://www.kaggle.com/datasets/bricevergnou/spotify-recommendation) and API (https://developer.spotify.com/documentation/web-api), since it had broad documentation and a variety of endpoints that related to audio features, streaming patterns, and artist trends. However, we found multiple issues with expired tokens, access limitations, and unpredictable endpoint responses on Spotify's API end. Therefore, we decided to make a change to our data in order to avoid these issues for our chatbot project. We ended up using data related to environmental and health in Canada, which some of us had used for our midterm. We combined a local CSV dataset on Canadian greenhouse gas emissions with a live API reporting provincial COVID-19 data from the Canada COVID-19 tracker platform. This allowed us to expand on the previous midterm project and have a more stable and informative chatbot to utilize historical and current data sources. Additionally, we thought this would be interesting to look more into as environmental issues are currently of prominence.

One of the biggest challenges faced in this project was refining the ETL process for the GHG (greenhouse gas emissions), as the dataset required cleaning and reshaping to handle the formatting, missing values, etc. The API also had separate issues since each province's record could vary slightly in schema. Integrating the processed data into a Flask application that could answer the user's questions was also a roadblock we faced multiple times. Connecting our Flask app to a SQLite database was fine, but deploying the chatbot to run from a Google Cloud Platform (GCP) instance was the main struggle. This included the firewall settings, public port exposure, and also getting the cURL commands to run within the GCP SSH. Once we got it to deploy, there were some issues with the chatbot logic and the formatting, which often failed to match user queries correctly. Based on how the query was written/formatted, it would read

another word within the query as the province in question. We were able to handle the errors to create a more robust and accurate response by changing our for-loop code slightly within the main.py. We also found that working remotely from different locations made communication more complex and less consistent than during the midterm project or other parts of the semester, as the lack of in-person interaction and finals reduced efficiency.

Through the process, we learned the importance of flexibility in project development. A key takeaway was understanding how to think quickly on our feet and adapt, as shown by changing our Spotify data to the environmental dataset. Since we already did a good amount on the Spotify data, it seemed like a waste not to use it. However, the change to the other dataset ultimately produced a more coherent and reliable product. We also gained deeper insight into structuring ETL pipelines and working with Flask's routing, error management mechanisms, and more. Through debugging live deployments and creating the Discord Chatbot, we also learned the value of clear logging and incremental testing throughout the process.

If we had more time, we would use more natural language processing tools like Transformers or spaCy to make our chatbot more conversational and able to understand different questions. We were also interested in adding a caching system so the bot wouldn't have to keep calling the API repeatedly; this would have helped with the efficiency. Another idea was to add visualizations, like charts showing emission levels or COVID trends, to make the information more manageable for the public to process.

In the end, even though we had to pivot a few times and ran into many technical issues, this project was helpful for learning how to work with APIs, clean data, build a Flask app, and deploy it to the cloud. Our final chatbot gives useful info about COVID-19 and emissions in Canada, and it's a great starting point if we ever want to add more features later on.