Exploring Optimizations in Anthophilia Taxonomic Research Using Retrieval-Augmented Generation

Team Members: Bennett Bishop, Daniel Yan, Casey Linden, Kasturi Sharma, Sean Reagan, Keon Dibley

Project Advisors: Katja Seltmann & Maddie Ostwald

Mentor: Professor Laura Baracaldo

University of California, Santa Barbara

Cheadle Center for Biodiversity & Ecological Restoration

Department of Statistics & Applied Probability Capstone Project

Individual Project Summary: Daniel Yan

Abstract:

A current problem that researchers involved in bee taxonomy face is the access and retrieval of relevant information for specific queries given the vast and often scattered nature of the data. In this project, our team developed a domain-specific chatbot, nicknamed *Chat G-Bee-T* that is designed to support taxonomic research at our advising organization, the Cheadle Center for Biodiversity & Ecological Restoration and beyond. In the process, we developed a data pipeline that would standardize the vastly different data sources we were provided and funneled it all into a vector database that enabled the chatbot to return grounded and accurate responses to user queries using retrieval augmented generation (RAG). Through evaluation by experts at the Cheadle Center, *Chat G-Bee-T* performed well in clarity and accuracy and while potential for improvement is present, the result of this project represents a significant improvement over baseline large language model (LLM) performance.

Personal Contribution:

The majority of my time in this project was dedicated to engineering and executing the data pipeline used to create embeddings for the vector store used in RAG retrieval. I wrote the script that processed the *cleaned-database.csv* file into natural language. Further, I wrote the scripts responsible for processing dichotomous keys found within our data into a format readable by the model. Finally, I was a main contributor to all presentation materials used throughout the project and also was responsible for maintaining the online repository and shared drive.

Supplementary Information:

I was the creator for all the scripts found in the *archives/archived_scripts* folder but specifically *cleaned-dataframe-analysis.ipynb* was responsible for the cleaning & conversion of the tabular data I was responsible for, *bees-of-the-world-key-trees.ipynb* which was the initial logic for a digital tree restructuring of the dichotomous keys. Furthermore, in the *scripts/keys* folder, *extraction_functions.py* is the compiled list of functions used in final key extraction that was run from *text_key_extraction.py* which created all the digital keys found in the *data/keys/nodes* folder and the *data/keys/terminal-nodes* folder. The final deliverable for key extraction that I was responsible for can be found in *data/texts/bees-of-the-world-key-descriptions.txt* while the tabulated conversion text is stored as *cleaned-database-text.txt* locally due to size constraints and is not found in the repository.