Exploring Optimizations in Anthophila Taxonomic Research Using Retrieval-Augmented Generation

Bennett Bishop[1], Daniel Yan[1], Keon Dibley[1], Kasturi Sharma[1], Casey Linden[1], Sean Reagan[1], Katja Seltmann[2], Maddie Ostwald[2]

[1] University of California, Santa Barbara; [2] Cheadle Center For Biodiversity & Ecological Restoration

UC SANTA BARBARA Cheadle Center for Biodiversity & Ecological Restoration

UC SANTA BARBARA | Data Science Initiative

Abstract

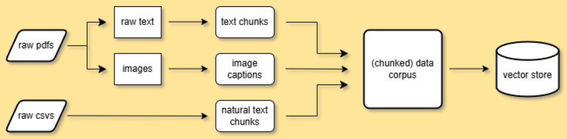
Although taxonomists have access to a massive amount of information, accessing this data is a major hassle for researchers and can be a barrier to entry for aspiring taxonomists. We created a chatbot for bee research that stores some of this information (provided by our advisors) and uses it to respond to user queries. Our approach was to implement a multimodal Retrieval Augmented Generation (RAG) pipeline, which matches user queries with relevant context in our bee data, returning answers based on related text chunks, along with images. This method yielded great results, with our advisors evaluating the chatbot highly compared to other prominent Large Language Models (LLMs). We are hosting the model as a website locally, and plan publish it in coordination with our advisors after more rigorous evaluation of its performance.

Introduction

In an effort to improve information retrieval for researches at the Big Bee Lab, our project takes a small subset of the data that researchers at the center commonly use and attempts to streamline the information retrieval process. The project was split into two major parts:

- Developing a **data pipeline** that would process multimodal data from a variety of sources & return it in formats that are optimal for use in our model
- Implementation of a **domain specific chatbot** that would retrieve and return the most accurate information based on the sources we provided

Data Processing Pipeline:

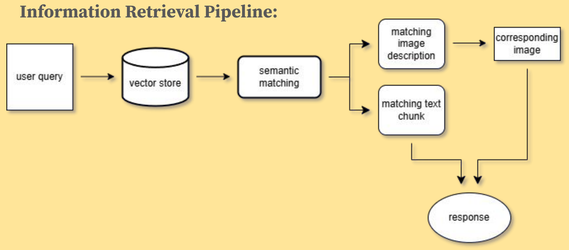


Problems with current retrieval methods:

- ***Traditional lookup*** is time consuming & inefficient
- ***Search engines*** do not information is often buried in academic papers or specific resources that are less accessible to the general public
- ***Large language models (LLMs)*** have poor performance with domain-specific responses such as in this case & will often hallucinate incorrect responses

Retrieval Augmented Generation (RAG):
Our project takes the resources provided by our advisors and implements RAG, a process that enhances the accuracy and depth of AI-generated answers by grounding them in real, retrieved data—rather than relying only on what the model was trained on.

Information Retrieval Pipeline:



Datasets

The subset of data that was presented for this project came in various forms and had to all be dealt with separately. It includes:

- ***PDFs***
  - Textbooks containing text, image, and dichotomous keys used for taxon classification
  - Contains vital information regarding bee structures and taxon characteristics used in identification methods
  - The three textbooks are: *Hymenoptera of the World*, *MMD*, and *The Bees of the World*
- ***CSV Files***
  - Contains collected bee sample occurrences, interactions, and taxonomic classifications
  - All tabulated data was converted into natural language
  - The csv files are: *discover-life*, *dory-bee-taxonomy*, *05-cleaned-database*, and *globi-bees-filtered*
- ***Raw Text Files***
  - Definitions of the anatomy of the bees and related species
  - The .txt file was hao.obo.txt

Below is a visualization of where bee specimens were collected across the world from our tabulated data.

Methodology

We designed the system architecture to be able to process multimodal data which would includes text, image, and tabulated data, taking care to focus on avoiding bias towards any certain type of data. A data pipeline was developed to standardize and clean the data our chatbot would use:

- ***Plain text*** was extracted from the PDFs, cleaned, chunked and embedded into the vector store
- ***Image data (diagrams & figures)*** was extracted from the PDFs and text descriptions of the images was generated through different methods:
  - When detailed captions near the figure in texts was present, we passed the image and description into the OpenAI API to generate the description
  - Otherwise, we scraped the raw PDF text for any references to the figure, passed all the scraped content into the OpenAI API with the image instead
- ***Dichotomous keys*** were converted into a digital tree structure & returned as a list of characteristics used in taxonomic identification and embedded as plain text
- ***Tabulated data*** was converted into natural language rowise and then embedded as plain text

Once the data processing was completed, we used the vector store we created from all the embeddings to augment OpenAI API responses to queries.

Chat G-Bee-T

The main result of our year of work is the chatbot that we have created and the database of cleaned data that we produced and uploaded to our vector store. The chatbot, named Chat G-Bee-T, is able to access that vector store and return them in a concise and effective manner in a carefully curated response that the user can then request elaboration or summary of if needed.

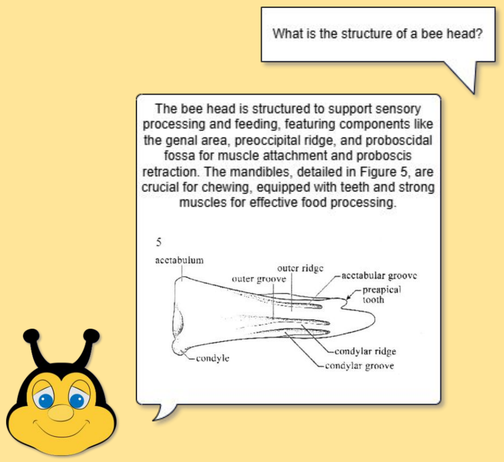
Our chatbot can respond with a combination of:

- Generated ***text responses*** based on our dataset
- ***Images*** that are relevant to the query
- ***Sources*** where the information was found
- ***Definitions*** of specific terminology used in the response

For each result used, our chatbot will grab the sources that it took the information from and includes them in a drop down menu. On top of this, it also grabs any key definitions from glossaries in the sources and adds them into another drop down to help with domain-specific terminology. Our chatbot performs best at responding to taxonomic questions. These questions can include:

- *Which bee family has a curved basal vein?*
- *How many green bees are there in California?*
- *What are key characteristics from the Apis family?*

Below is an example question and output from our chatbot shown as a simplified graphic. (not shown: sources and definitions)



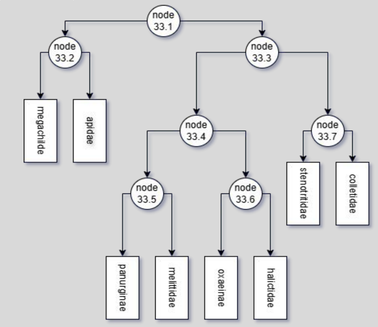
Performance Evaluation:
With the help from our advisors, we created a dataset of around 20 plausible questions that could be asked by a user. We then asked our chatbot these questions and asked the advisors to grade the response on three different criteria: correctness, clarity, and completeness. These were graded on a scale of 1 - 5 and then averaged to figure out where we needed to improve. Clarity was our best scoring category while correctness and completeness lagged behind. We decided to update our prompt engineering in order to improve these two lagging categories.

Room For Further Improvement:
It is not the best at responding to areas outside of its expertise. These include bee behavior and anything else that is not present in the subset of bee data that we received. In this case, it will do its best to notify the user that it does not have adequate information to formulate a proper response.

Dichotomous Keys as a Tree Structure

As a part of our data pipeline, a large effort was made to translate written dichotomous keys used for taxon identification from written key structure into a format that would be readable as plain text. The keys were extracted from the PDF text, cleaned and then formatted into tabular structure that expressed the keys as a tree structure.

Below is a simplified visualization of a dichotomous key for family identification. (not shown: decision conditions at each node)



Future Work

Current Use Cases:

- ***Streamlines Discovery:*** faster access to required information, speeds up species identification & recognition, and streamlines access to visuals and metadata
- ***Democratizes Knowledge:*** makes complex bee taxonomy accessible and understandable to researchers, students, and enthusiasts alike

Potential Extensions:

- ***Improved Evaluation:*** systematic testing of chatbot accuracy and reliability
- ***Prompt Optimization:*** fine-tuning prompts for clearer, more precise answers
- ***Visual & Interactive Dichotomous Keys:*** incorporating the digitized keys into a user-interactive format for improved taxon identification
- ***Expansion of Database:*** expand the chatbot knowledge base with more resources on a broader range of relevant domain knowledge

References and Acknowledgments

Goulet, H., & Huber, J. T. (1993). *Hymenoptera of the world: An identification guide to families*. Centre for Land and Biological Resources Research.

Michener, C. D., Danforth, B. N., & MacGinley, R. J. (1994). *The bee genera of North and Central America: (hymenoptera: Apoidea)*. Smithsonian Institution Press.

Michener, Charles D. (2007). *The bees of the world*. Johns Hopkins University Press.

We would like to thank the Cheadle Center for Biodiversity and Ecological Restoration and the UCSB Department of Probability and Statistics for all the help and support throughout the year. Without whom, this project would not have been possible.