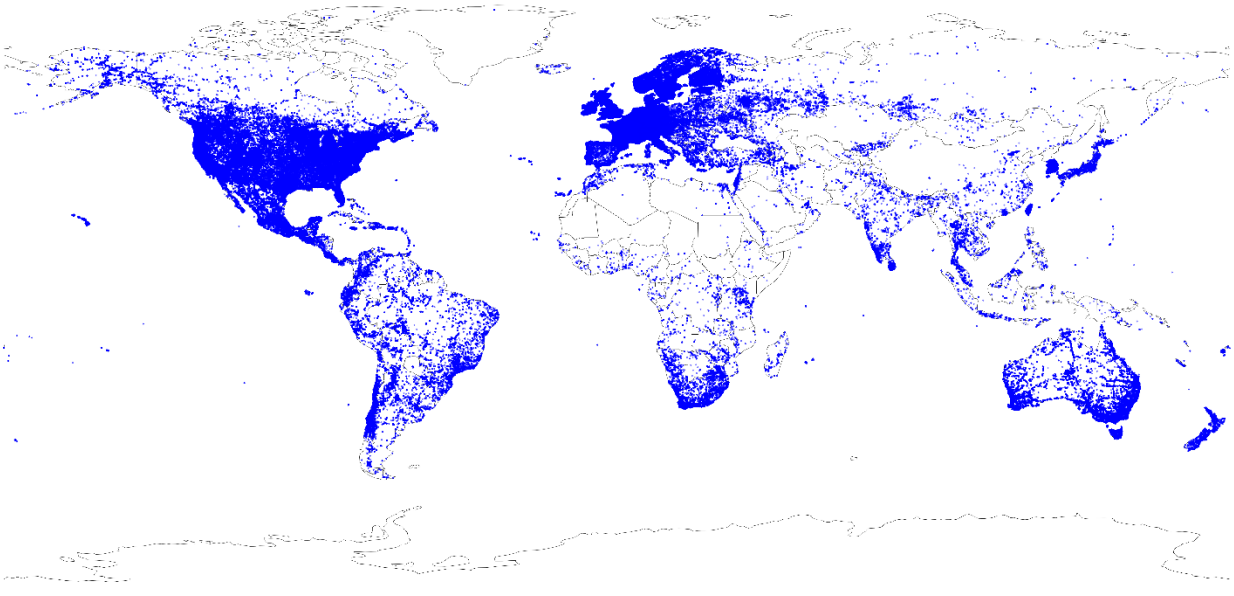
Cheadle Center for Biodiversity & Ecological Restoration

Abstract

With over 20,000 species of bees worldwide and a wealth of complex and specific etomological data & information, researchers and conservationists face challenges in retrieving specific information they seek. Additionally, general purpose Large Language Models (LLMs) currently lack the specific domain knowledge required to answer specific questions in specialized fields. This project seeks to improve upon current LLMs in this specific field by building a multimodal retrival augmented generation (RAG) pipeline that can handle text, image and tabulated data.



"According to all known laws of aviation, there's no way a bee should be able to fly… The bee, of course, flies anyway – because bees don't care about what humans think is impossible" – Bee Movie (2007)

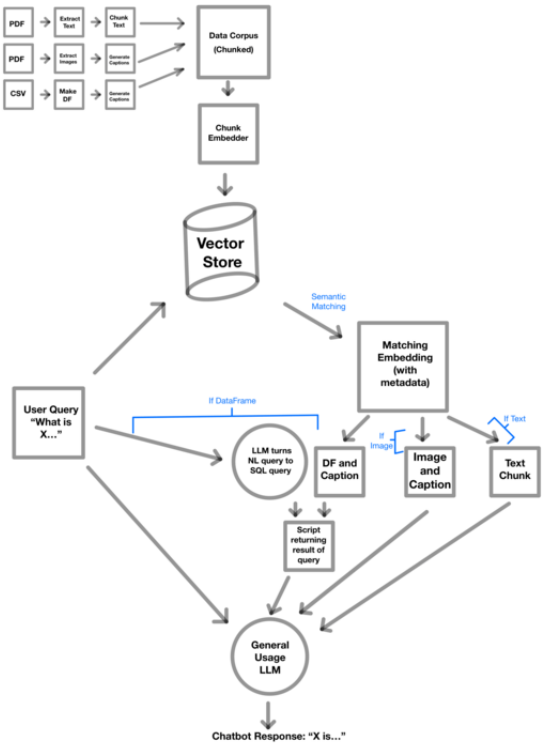
Data Overview

- PDFs containing text, images/diagrams & species identification keys
- Tabular data on species, scientific names & general information on each species
- Database of individual bee specimens collected

Current Progress

- Base RAG model with information from PDF data sources
  - Returns information from data sources related to query
  - Image retrieval for relevant queries
- Data extraction & cleaning for RAG model:
  - Convert tabular data into natural language
  - Image & caption extraction from PDFs
- Evaluation metrics for LLM model – synthetic evaluation dataset
  - Evaluate model ability to extract correct context

System Architecture



Future Plans

- Complete caption & image embedding
- Convert identification keys into data structure
- Fine tune RAG model with evaluation metrics
- Website & front-end development

ChatGBeeT

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

Project Sponsors:
UCSB Cheadle Center for Biodiversity & Ecological Restoration: Katja Seltmann, Maddie Ostwald

Project Advisor:
Professor Laura Baracaldo