# Cheadle Center for Biodiversity and Ecological Restoration

## ChatGBeeT

**Presented by: Bennett Bishop, Casey Linden, Daniel Yan, Kasturi Sharma, Keon Dibley, and Sean Reagan**

**Sponsors: Katja Seltmann and Maddie Ostwald**

# Hand Out QR Code

**01** The Big Bee lab has research dating back decades. Very thorough but difficult to go through quickly.

**02** Utilizing an LLM (large language model) to facilitate searching through the extensive data available

**03** Our end goal is to build a Multimodal RAG (Retrieval Augmented Generation) Pipeline with Text, Image and Tabulated data and a website to host it

# Overview

We are building a large-language model (LLM) for our sponsors, the Cheadle Center for Biodiversity and Ecological Restoration.

# What is RAG?

- Pre-trained LLMs are trained on an initial data corpus, but lack access to up-to-date, or niche data.
- Retrieval-Augmented Generation (RAG) enhances LLMs by allowing access to a knowledge base before generating responses.
- Ensuring answers are grounded in external data rather than relying solely on the model's pre-training.
- Useful for domain-specific knowledge, like legal, medical, or technical fields, where up-to-date or proprietary data is essential.

# Data Overview

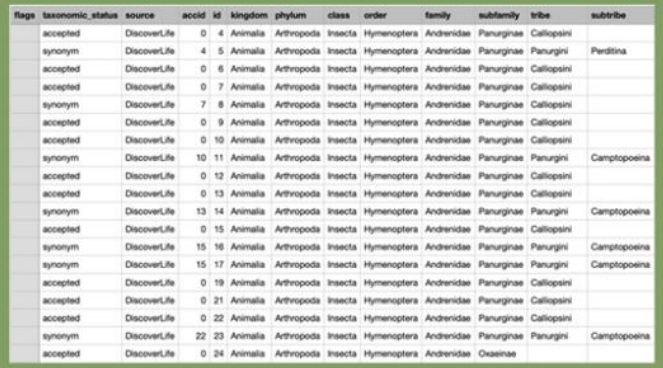
1. MMD
2. Bees of the World
3. 05_cleaned_database
4. Dory-bee-taxonomy
5. Discover life
6. Globi-bees-filtered
7. Hymenoptera of the World
8. Hao.obo

# Data Formats/Challenges

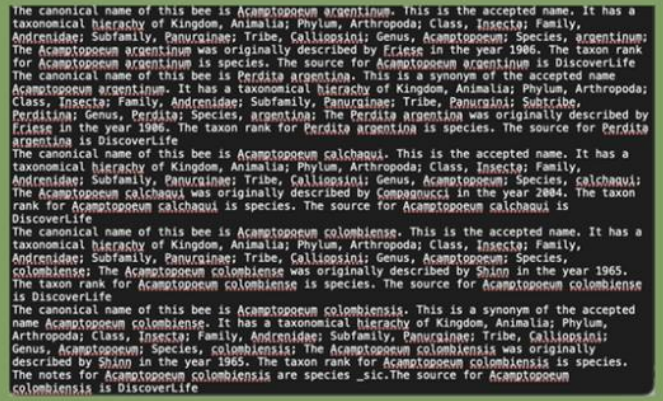
- PDFs:
  - Three books contain information on bees' structure, identification, and taxonomy.
  - Text data, image data, decision tree-like key data.
- CSVs:
  - Four CSV files of global bee occurrences, interactions, and taxonomic classifications.
  - Each row must be converted to natural language.
- TXT
  - One TXT file with definitions of the anatomy of bees and related species.

# Tabular Data

- Converted tabular data into natural language
- Used the globi, dory-bee, discover life, and 05_cleaned datasets
- Turned them from csv format to txt
- Having them be in natural language to facilitate feeding them through LLM

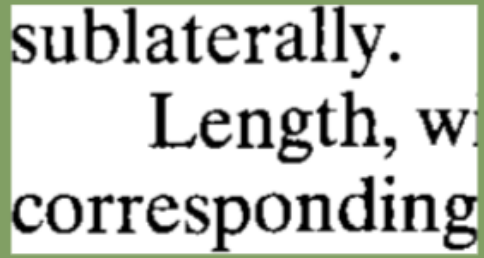


| flags | taxonomic_status | source | accid | id | kingdom | phylum | class | order | family | subfamily | tribe | subtribe |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | accepted | DiscoverLife | 0 | 4 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | synonym | DiscoverLife | 4 | 5 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Panurgini | Perditina |
| | accepted | DiscoverLife | 0 | 6 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | accepted | DiscoverLife | 0 | 7 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | synonym | DiscoverLife | 7 | 8 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | accepted | DiscoverLife | 0 | 9 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | accepted | DiscoverLife | 0 | 10 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | synonym | DiscoverLife | 10 | 11 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Panurgini | Camptopoeina |
| | accepted | DiscoverLife | 0 | 12 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | accepted | DiscoverLife | 0 | 13 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | synonym | DiscoverLife | 13 | 14 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Panurgini | Camptopoeina |
| | accepted | DiscoverLife | 0 | 15 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | synonym | DiscoverLife | 15 | 16 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Panurgini | Camptopoeina |
| | synonym | DiscoverLife | 15 | 17 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Panurgini | Camptopoeina |
| | accepted | DiscoverLife | 0 | 19 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | accepted | DiscoverLife | 0 | 21 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | accepted | DiscoverLife | 0 | 22 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Calliopsini | |
| | synonym | DiscoverLife | 22 | 23 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Panurginae | Panurgini | Camptopoeina |
| | accepted | DiscoverLife | 0 | 24 | Animalia | Arthropoda | Insecta | Hymenoptera | Andrenidae | Oxaeinae | | |

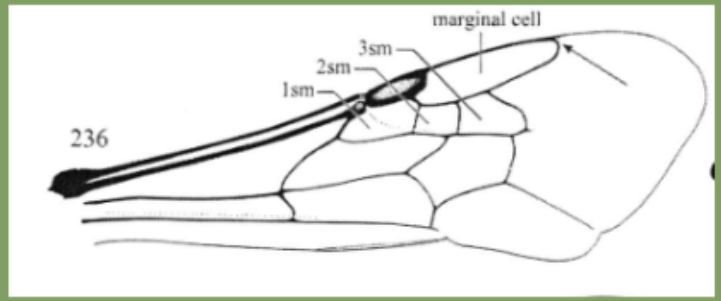


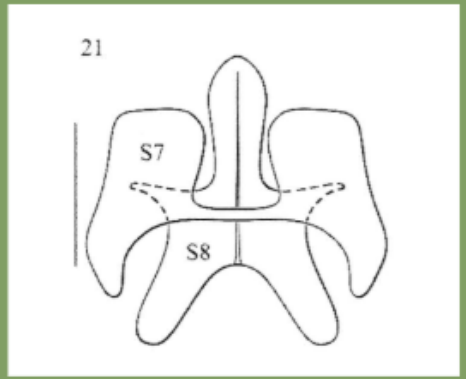
The canonical name of this bee is Acamptopoeum argentinum. This is the accepted name. It has a taxonomical hierachy of Kingdom, Animalia; Phylum, Arthropoda; Class, Insecta; Family, Andrenidae; Subfamily, Panurginae; Tribe, Calliopsini; Genus, Acamptopoeum; Species, argentinum; The Acamptopoeum argentinum was originally described by Friese in the year 1906. The taxon rank for Acamptopoeum argentinum is species. The source for Acamptopoeum argentinum is DiscoverLife
The canonical name of this bee is Perdita argentina. This is a synonym of the accepted name Acamptopoeum argentinum. It has a taxonomical hierachy of Kingdom, Animalia; Phylum, Arthropoda; Class, Insecta; Family, Andrenidae; Subfamily, Panurginae; Tribe, Panurgini; Subtribe, Perditina; Genus, Perdita; Species, argentina; The Perdita argentina was originally described by Friese in the year 1906. The taxon rank for Perdita argentina is species. The source for Perdita argentina is DiscoverLife
The canonical name of this bee is Acamptopoeum calchaqui. This is the accepted name. It has a taxonomical hierachy of Kingdom, Animalia; Phylum, Arthropoda; Class, Insecta; Family, Andrenidae; Subfamily, Panurginae; Tribe, Calliopsini; Genus, Acamptopoeum; Species, calchaqui; The Acamptopoeum calchaqui was originally described by Compagnucci in the year 2004. The taxon rank for Acamptopoeum calchaqui is species. The source for Acamptopoeum calchaqui is DiscoverLife
The canonical name of this bee is Acamptopoeum colombiense. This is the accepted name. It has a taxonomical hierachy of Kingdom, Animalia; Phylum, Arthropoda; Class, Insecta; Family, Andrenidae; Subfamily, Panurginae; Tribe, Calliopsini; Genus, Acamptopoeum; Species, colombiense; The Acamptopoeum colombiense was originally described by Shinn in the year 1965. The taxon rank for Acamptopoeum colombiense is species. The source for Acamptopoeum colombiense is DiscoverLife
The canonical name of this bee is Acamptopoeum colombiensis. This is a synonym of the accepted name Acamptopoeum colombiense. It has a taxonomical hierachy of Kingdom, Animalia; Phylum, Arthropoda; Class, Insecta; Family, Andrenidae; Subfamily, Panurginae; Tribe, Calliopsini; Genus, Acamptopoeum; Species, colombiensis; The Acamptopoeum colombiensis was originally described by Shinn in the year 1965. The taxon rank for Acamptopoeum colombiensis is species. The notes for Acamptopoeum colombiensis are species _sic.The source for Acamptopoeum colombiensis is DiscoverLife

# Getting the Images




- Goal: Extract images from the PDFs provided, store them in a cloud warehouse, embed them for semantic search
- Initially, we wrote a python script to extract images, but it did not work
- Current status: we manually extracted images for one of the pdfs

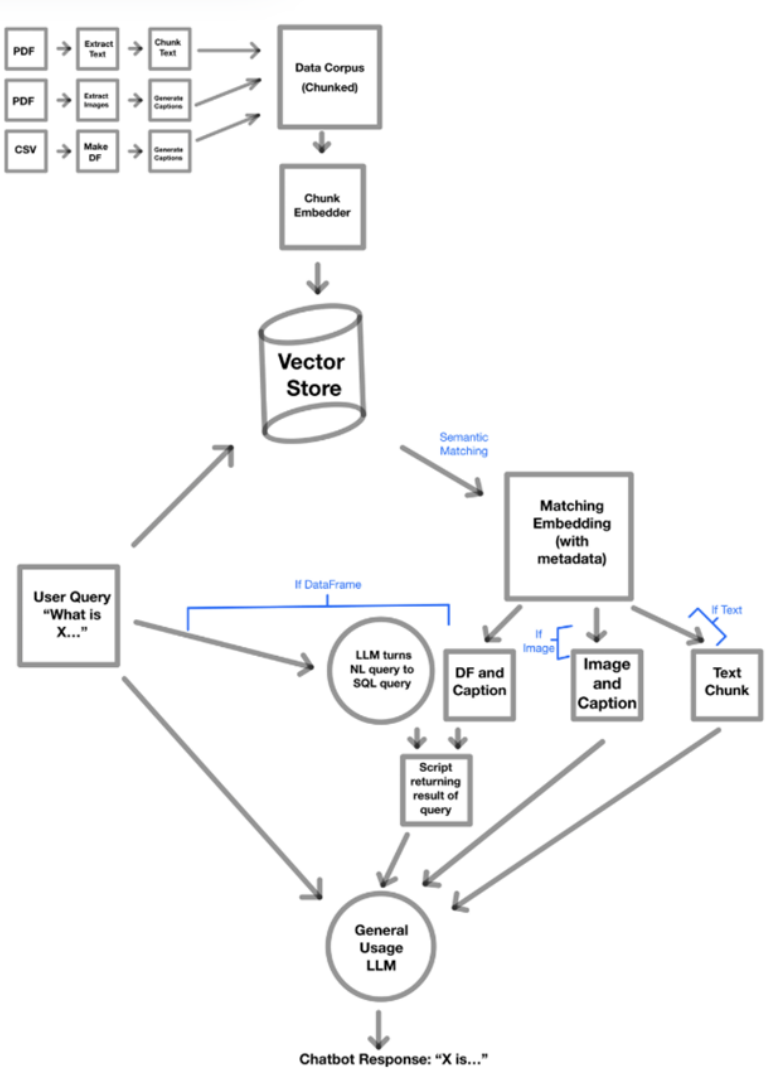


# Image Caption Generation / Extraction

- We first tried CLIP, but our figures are so niche, it did not work!
- So, why not use pre-existing written references to the figures, and generate our own captions?
- Figure descriptions are scattered throughout the PDFs, so we wrote a script to extract all references to each figure. We then fed these into an LLM to generate descriptions of the figures.
- These image descriptions are embedded into our vector store (with the actual image url in the metadata) alongside our text data, so when a user makes a query, both text and image data is returned!

# System Architecture



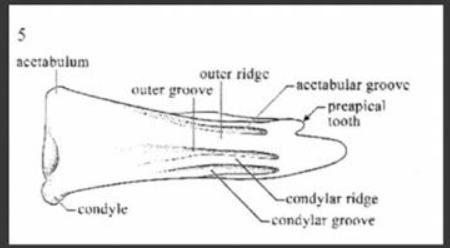
# Getting Context:

```
[ ] # Now, let's see if we can retrieve the right embeddings
    chroma_client = chromadb.PersistentClient(path="./chroma_db")  # Store data persistently
    collection_name = 'neww_collection'
    question = 'What are mandible pairs?'
    results = get_relevant_docs(question, chroma_client, collection_name)

    print(f"Retrieved IDs: {results['ids']}")
    print(f"Retrieved Documents: {results['documents']}")
```

```
Retrieved IDs: [['MMD-Figures/page_16_img_5.png', 'MMD_CHUNK_169', 'MMD_CHUNK_156']]
Retrieved Documents: [["**Figure Description of Fig. 5: Bee Mandibular Structures**\n\nFig. 5 illustrates the detailed anato
```

# Retrieving Results:

```
[ ] import textwrap
    # Example Usage
    question = "What are mandible pairs?"
    retrieval_results = get_relevant_docs(question, chroma_client, collection_name)
    answer = generate_answer_with_images(OpenAI(api_key=OPENAI_API_KEY), question, retrieval_results)

    print(answer)  # Displays the OpenAI-generated response
```




Mandible pairs refer to the two mandibles found in some insects, such as bees, which are used primarily for grasping, biting, cutting, or crushing food. In the context of bees, as illustrated in Figure 5, the mandibles are detailed anatomical structures that include components like preapical teeth and a basal part. These structures enable bees to effectively process food, contributing to their feeding habits. The diagram in Figure 5 provides a close-up view of these mandibular components, highlighting their importance in the bee's anatomy and feeding mechanisms.

# Evaluation Metrics

1. Developed a synthetic **evaluation dataset** of question-answer pairs using multiple LLM agents from Hugging Face
2. Assessed model performance by evaluating its ability to **extract the correct context** for answering questions
3. Optimized model parameters to **maximize accuracy** in answering questions based on the evaluation dataset

| | | |
|---|---|---|
| sternum and, with a hooked needle, pull out the genitalia and hidden stema. In most cases such dissection is not difficult, but in the Megachilidae the numerous hidden stema are firmly connected to one another and to the terga laterally and are often delicate medially, so that successful dissection may be difficult. Beginners should start with other groups. | What family of bees has numerous hidden sterna? | Megachilidae |

# Current Progress

- Converting tabular data into natural language
- Extracting images from PDFs
- Extracting captions from PDFs
- Generating captions
- Determining evaluation metrics

# Future Plans

- Focusing on caption and image embedding
- Embedding text and tabular data into vector space
- Convert identification keys into data structure format
- Fine-tuning RAG model
- Continuous testing with the sponsors
- Developing a website to host the chatbot

# Bonus: CCBER Exhibit!


# Thank you!