Integrating Large Language Models and the iDigBio Portal for Conversational Data Exploration and Retrieval

Michael J Elliott[‡], Manuel Luciano[‡], Jose Fortes[‡]

‡ University of Florida, Gainesville, United States of America

Corresponding author: Michael J Elliott (mielliott@ufl.edu)

Abstract

The advent of cloud-based large language model (LLM) services such as ChatGPT (Generative Pre-Trained Transformer) has given rise to a wide array of novel artificial intelligence (AI) applications. In particular, LLMs have been used to power AI assistants that serve as intermediaries between human users and online web services, namely, web-based application programming interfaces (web APIs). These AI assistants allow users to make requests in natural language to initiate complex processes, ranging from searching a database to making a reservation.

We are exploring the development of AI assistants that can intelligently search for and process species occurrence data served by the Integrated Digitized Biocollections (iDigBio) Portal. Though the portal already provides a human-friendly search interface, it is tailored for a very particular use case: finding and inspecting records that match the user's search parameters. However, the underlying iDigBio APIs that power the search interface offer direct access to biodiversity data and metadata that can support a wider range of applications. An LLM-powered AI assistant with access to such APIs has the potential to redefine how researchers discover and interact with scientific data by 1) allowing users to interact with scientific databases using natural language, 2) serving as a single unifying interface for many different use cases, and 3) enhancing the user's experience with AI insights that are backed by citable, curated data.

Fig. 1 demonstrates a prototype chatbot we have developed, which interfaces with the iDigBio Portal. The chatbot uses OpenAI's GPT-4 to understand user requests and call the appropriate APIs as needed. It currently has access to the following iDigBio APIs:

- The **Search API** allows the chatbot to perform the same search functions as the portal search interface. The results of the search can be observed by either directly calling the API from the user's web browser, opening the existing portal search interface with generated search parameters, or visualizing the geographic distribution of matched records on an interactive map.

- The **Download API** allows the chatbot to package search results as a Darwin Core Archive to be delivered by email.
- The **Summary API** allows the chatbot to collect statistics across all of iDigBio’s data, such as a breakdown of record counts for unique species reported within a particular geographic region.

Because the chatbot is intended for use by researchers, transparency is critical. When responding to user requests, LLMs often include their own internalized knowledge—which may be unreliable but difficult to verify—or make up information entirely. Thus, it must be abundantly clear how the chatbot forms its responses, in particular how the LLM interprets user requests and how it queries external APIs, such that users may independently assess the correctness of the chatbot's actions and link its conclusions back to data. The approach we adopted for the design of our prototype is illustrated in Fig. 2; whenever the chatbot initiates a data retrieval or analysis process, a detailed log of that process is embedded in the conversation, including the sequence of actions taken and their inputs and outputs.

The chatbot's behavior is tightly controlled by the rigid use of *specialized AI agents* with *expert-defined validators* (Fig. 3). Importantly, we do not allow the LLM to respond freely to user requests; rather, we use a fixed set of AI agents that are specialized to produce structured outputs that follow strict schema as enforced by the agent's assigned validator. An *orchestrator* agent is responsible for deciding which agent is most appropriate to handle each request. If no agent is chosen, the system does not attempt to answer the user's request.

As a prototype, the functionality of the chatbot is currently limited to the few illustrative use cases we have outlined. However, as the system is incrementally refined and expanded, we envision the single chatbot interface to be of interest to both the general public and researchers alike. For the general public, it may be a useful tool to learn about biodiversity in their local community and around the world. Meanwhile, researchers may find the chatbot useful for quickly navigating and exploring iDigBio's hosted data and APIs. The prototype is hosted online at chat.acis.ufl.edu with source code in GitHub.

Keywords

species occurrence records, LLM, Artificial Intelligence (AI), chatbot

Presenting author

Michael Elliott

Presented at

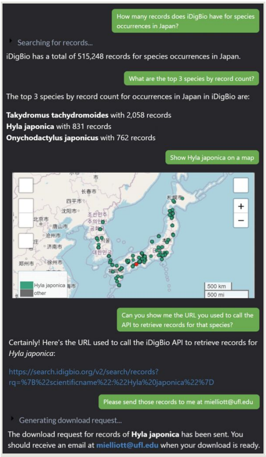
SPNHC-TDWG 2024

Acknowledgements

The research reported in this work was funded in part by grants from the National Science Foundation (DBI 2027654) and the AT&T Foundation.

Conflicts of interest

The authors have declared that no competing interests exist.



How many records does iDigBio have for species occurrences in Japan?

Searching for records...

iDigBio has a total of 515,248 records for species occurrences in Japan.

What are the top 3 species by record count?

The top 3 species by record count for occurrences in Japan in iDigBio are:

Takydromus tachydromoides with 2,058 records
Hyla japonica with 831 records
Onychodactylus japonicus with 762 records

Show Hyla japonica on a map

Can you show me the URL you used to call the API to retrieve records for that species?

Certainly! Here's the URL used to call the iDigBio API to retrieve records for *Hyla japonica*:

https://search.idigbio.org/v2/search/records?rq=%7B%22scientificname%22:%22Hyla%20japonica%22%7D

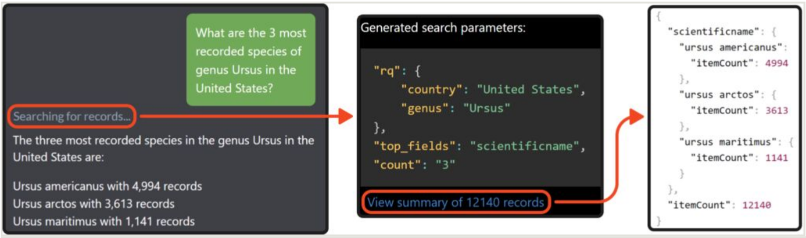
Please send those records to me at mielliott@ufl.edu

Generating download request...

The download request for records of **Hyla japonica** has been sent. You should receive an email at mielliott@ufl.edu when your download is ready.

Figure 1.

An example conversation demonstrating our prototype chatbot's ability to perform record searches, count records, visualize species occurrences on a map, and initiate download requests.



What are the 3 most recorded species of genus Ursus in the United States?

Searching for records...

The three most recorded species in the genus Ursus in the United States are:

Ursus americanus with 4,994 records
Ursus arctos with 3,613 records
Ursus maritimus with 1,141 records

Generated search parameters:

```
"rq": {
    "country": "United States",
    "genus": "Ursus"
},
"top_fields": "scientificname",
"count": "3"
```

View summary of 12140 records

```
{
  "scientificname": {
    "ursus americanus":
      "itemCount": 4994
    },
    "ursus arctos": {
      "itemCount": 3613
    },
    "ursus maritimus": {
      "itemCount": 1141
    }
  },
  "itemCount": 12140
}
```

Figure 2.

Gray text (left) in the chatbot’s responses can be expanded to reveal information about actions it initiates. In this example, this includes a generated query to the iDigBio Summary API (middle) and a link to view the record counts returned by the API (right).

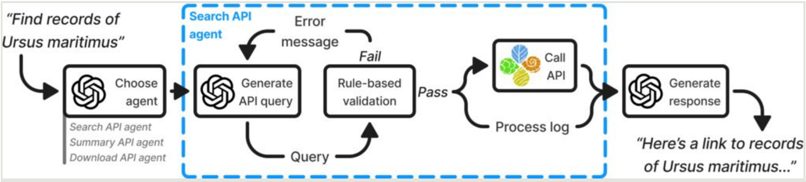


Figure 3.

Our prototype chatbot makes use of LLM-powered agents paired with expert-defined validators. The OpenAI logo indicates a GPT-4-powered processes. The iDigBio logo indicates processes that call the iDigBio APIs.