BotGBIF: Una herramienta para consultar datos de GBIF en lenguaje natural

Marcos Zárate[1,2], Gustavo Nuñez[1,2], Dario Ceballos[1,3], Macarena Repetto[2] and Nora Morgan[4]

[1] Centro para el Estudio de Sistema Marinos (CESIMAR-CONICET)
zarate@cenpat-conicet.gob.ar
[2] Laboratorio de Investigaciones en Informática (LINVI-UNPSJB)
gnunez@cenpat-conicet.gob.ar
[3] Facultad de Tecnología Informática, Universidad Abierta Interamericana (FTI-UAI)
dceballos@cenpat-conicet.gob.ar
[4] Facultad de Arquitectura, Diseño y Urbanismo, Universidad de Buenos Aires (FADU-UBA)
noraemorgan@gmail.com

Abstract. BotGBIF es un prototipo de chatbot que facilita el acceso a datos de biodiversidad del Sistema Global de Información sobre Biodiversidad (GBIF) mediante consultas en lenguaje natural. Utiliza Streamlit para la interfaz, la API REST de GBIF para obtener datos y GPT-4 como modelo de lenguaje. Su objetivo es simplificar la interacción con los datos para usuarios sin conocimientos técnicos, permitiendo consultas intuitivas y accesibles para investigadores, conservacionistas y el público general.

Keywords: Modelo Extenso de Lenguaje, GPT-4, API rest, GBIF

BotGBIF: A tool to query GBIF data in natural language

Abstract. This project introduces BotGBIF, a prototype chatbot designed to enhance user access to biodiversity data from the Global Biodiversity Information Facility (GBIF) through natural language queries. Utilizing the Python Streamlit library for the frontend, the GBIF REST API for data retrieval, and the capabilities of GPT-4 as a large language model (LLM), BotGBIF aims to simplify the interaction with complex datasets for users who may lack technical expertise in utilizing the GBIF portal. By integrating the GBIF API with GPT-4, the chatbot allows users to ask questions in everyday language and receive intuitive responses, thereby making biodiversity data more accessible to a range of stakeholders, including researchers, conservationists, and the general public.

Keywords: Large Language Model, GPT4, API rest, GBIF

1 Introduction

Biodiversity is essential for maintaining ecological balance and ensuring the survival of various species, including humans (Eldredge, 2000). Access to accurate and up-to-date biodiversity data is critical for scientific research, informed policymaking, and effective conservation efforts (Ladle and Whittaker, 2020). The Global Biodiversity Information Facility (GBIF)[5] provides an extensive repository of biodiversity data collected from diverse sources worldwide. Despite its substantial value, users, especially those without programming knowledge or familiarity with GBIF's complex interface, often face significant challenges when attempting to retrieve and utilize this wealth of information.

Recent advancements in Natural Language Processing (NLP) through the development of LLMs have shown remarkable capabilities in interpreting and generating human-like text from both structured and unstructured data (Brown et al., 2020; Devlin et al., 2019). These models can analyze extensive datasets, extract meaningful patterns, and provide coherent responses, making them valuable tools for applications in biodiversity informatics. The rapid growth of open-source LLMs since the release of ChatGPT (OpenAI, 2025) has stimulated the creation of various domain-specific chatbots that enhance user interaction with complex datasets (Touvron et al., 2023).

RESTful APIs have become a standardized method for exposing functionalities and datasets in scientific domains (Fielding, 2000). GBIF offers a RESTful API[6] that enables programmatic access to its rich biodiversity records; however, successfully querying these APIs usually requires technical proficiency, including knowledge of endpoints, authentication mechanisms, and the construction of query parameters. Consequently, many potential users, such as researchers, educators, and policymakers, may miss out on the valuable data GBIF provides.

To address this challenge, we propose BotGBIF, an innovative chatbot that integrates LLMs with the GBIF API, allowing users to retrieve biodiversity information through natural language queries. This initiative seeks to improve the accessibility and usability for individuals without technical backgrounds, thereby facilitating broader engagement with GBIF's extensive database. By leveraging an LLM capable of dynamically interacting with APIs, BotGBIF aims to enhance the user experience in biodiversity data retrieval, potentially enriching results with supplementary information from other relevant knowledge sources.

This research is structured around research questions (RQs) regarding the feasibility and impact of integrating LLMs within biodiversity data environments:

- RQ1: Can natural language questions be effectively answered using only the GBIF API, without relying on external data sources?
- RQ2: Is it possible for an LLM to intelligently utilize additional services to supplement information retrieved from GBIF, ensuring more comprehensive responses?

[5] https://www.gbif.org/
[6] https://techdocs.gbif.org/en/openapi/

– RQ3: How can the integration of LLMs enhance accessibility and usability for non-technical users, promoting wider adoption and more informed decision-making?

This research contributes to the growing intersection of artificial intelligence and biodiversity informatics, showcasing how AI-driven tools can democratize access to complex scientific datasets and foster data-driven insights in ecological and conservation efforts.

The remainder of this paper is organized as follows: Section 2, provides a detailed overview of the relevant literature related to biodiversity informatics and natural language processing technologies. Section 3, outlines the methodology employed in developing BotGBIF, including technical specifications and implementation strategies. Section 4, shows the development process used for BotGBIF. Section 5, presents the case study that demonstrates the usefulness of the application. Finally, Section 6 and Section 7 discusses the implications of these findings, offers conclusions, and suggests potential avenues for future research

2 Related work

The intersection of LLMs and scientific data retrieval has emerged as a critical area of research, with numerous attempts to bridge the gap between complex scientific databases and user-friendly interfaces (Nejjar et al., 2025; Vert, 2023). Previous research has highlighted the challenges of translating natural language queries into precise scientific data retrievals, particularly in specialized domains like biodiversity informatics (Domazetoski, 2024). Scholars have explored various approaches to enhancing data accessibility, with early efforts focusing on developing domain-specific search interfaces that require significant technical expertise. The advent of advanced language models has opened new possibilities for more intuitive data interaction. Researchers like (Chen et al., 2022) demonstrated the potential of AI-driven approaches in scientific data retrieval, showing how natural language processing could transform complex API interactions into more accessible knowledge exploration.

In the realm of biodiversity informatics, multiple challenges have persisted in making large-scale biodiversity databases more approachable for non-technical users. Previous systems typically required users to have extensive knowledge of specific query languages or complex interface designs. The work of (von Wettberg and Khoury, 2022) highlighted the significant barriers that exist in current biodiversity data retrieval systems, emphasizing the need for more user-friendly approaches that can democratize access to scientific information.

One significant effort in this area is GBIF's IPT (Integrated Publishing Toolkit), which facilitates biodiversity data publication. The IPT supports the transformation of raw biodiversity data into standardized formats for global sharing. However, despite its widespread use, interacting with the IPT often requires technical knowledge, making it less accessible for users unfamiliar with data publishing protocols (Robertson et al., 2014).

Another notable initiative is BioGPT, an LLM developed specifically for biomedical and biological data processing. Unlike general-purpose language models, BioGPT is fine-tuned on life sciences literature, enabling it to extract meaningful insights from complex datasets (Luo et al., 2022). Although BioGPT focuses more on textual data rather than structured biodiversity repositories, it demonstrates the potential of domain-specific LLMs to improve accessibility to scientific data.

More recently, GBIF Norway developed ChatIPT (Norway, 2024), a chatbot assistant that guides data holders in transforming their spreadsheets into GBIF-ready datasets. ChatIPT helps users navigate routine data publishing questions and processes their submissions efficiently. This tool won the 2024 Ebbe Nielsen Challenge, recognizing its contribution to improving biodiversity data accessibility.

Additionally, the Conservation CoPilot (Cambridge, 2024), developed by the University of Cambridge, applies AI-driven decision support for biodiversity conservation projects. By leveraging large-scale computing resources, this tool assists conservationists in identifying optimal strategies to mitigate biodiversity loss.

In the context of biocollections, iDigBio has introduced a conversational interface to facilitate access to its biological collection records (iDigBio, 2024). The chatbot allows users to interact with biodiversity occurrence data, metadata, and references from the Biodiversity Heritage Library using natural language queries.

The emergence of LLMs has provided a promising avenue for addressing these accessibility challenges. Preliminary studies by Wang et al. explored the potential of AI-mediated data retrieval, demonstrating how sophisticated language models could interpret complex user queries and translate them into structured data requests (Wang et al., 2023). However, these early approaches often struggled with domain-specific nuances and the precise translation of scientific queries.

BotGBIF builds upon these foundational works by introducing a novel approach that addresses key limitations in existing systems. Unlike previous attempts, the proposed system implements a refined methodology for query generation, incorporating user-guided parameter selection and dynamic API interaction. This approach mitigates common challenges such as query ambiguity and provides a more reliable mechanism for retrieving biodiversity data through natural language interfaces. The research contributes to a growing body of work that seeks to make scientific databases more accessible and user-friendly. By leveraging the capabilities of GPT-4 and implementing a structured approach to API querying, BotGBIF represents a significant step forward in bridging the gap between complex scientific data repositories and end-users with varying levels of technical expertise. The system not only facilitates easier access to biodiversity information but also provides a framework for future developments in AI-mediated scientific data retrieval.

3 Methods

The development of BotGBIF followed an iterative and exploratory approach, combining techniques from software engineering, artificial intelligence, and biodiversity informatics. The system integrates a LLM with the GBIF API, enabling users to retrieve biodiversity data and metadata through a user-friendly conversational interface. The methodology is structured into the following key phases:

System Architecture and Design

To facilitate seamless interaction between users and biodiversity data, BotGBIF was designed as a modular system consisting of three main components:

- **User Interface (UI):** A web-based interface developed with Streamlit[7], allowing users to input queries, configure search parameters, and visualize results dynamically.
- **API Query Module:** This component processes user queries, constructs API calls, and retrieves responses from GBIF's API.
- **Natural Language Processing (NLP) Module:** An LLM (GPT-4) interprets user queries, extracts relevant parameters, and reformulates API responses into natural language explanations.

The modular approach ensures scalability, flexibility, and interoperability with other biodiversity data services. Figure 1 illustrates the system architecture of BotGBIF.

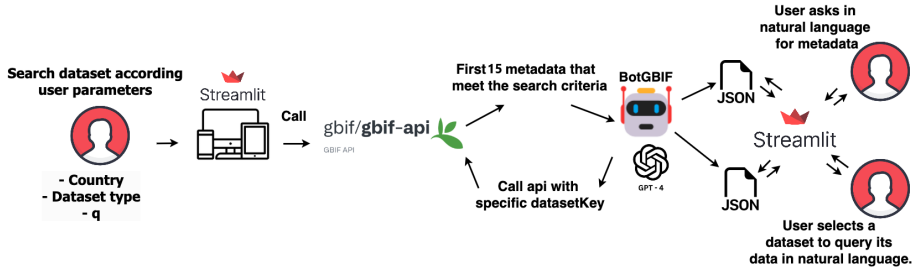

Fig. 1. Illustrates the workflow of BotGBIF, from user query input to retrieving and processing biodiversity data. Users specify search parameters (country, dataset type, and keyword) via a Streamlit interface, which queries the GBIF API. The system retrieves relevant dataset metadata and allows users to interact with it using natural language through GPT-4. The chatbot facilitates querying specific datasets, providing structured responses in JSON format.

[7] https://streamlit.io/

GBIF API Exploration and Query Construction

A comprehensive exploration of the GBIF API was conducted to identify the available endpoints, query parameters, and data structures. Initial tests involved manually querying the API to determine the best strategies for data retrieval and filtering. Key findings from this exploration included:

- The **occurrence** and **dataset search** endpoints were the most relevant for user queries.
- Pagination mechanisms were necessary to handle large query results efficiently.
- API responses were returned in JSON format, requiring post-processing for readability.

Early tests revealed that LLMs could not directly interpret raw API responses effectively. Consequently, the system was designed to generate API calls dynamically based on user-selected filters, such as **country, dataset type, and keywords**.

Natural Language Query Processing

BotGBIF leverages GPT-4 to process user queries and interact with biodiversity datasets in a conversational manner. The query workflow consists of three stages:

1. **User Input Interpretation:** The model identifies relevant keywords and search parameters.
2. **API Call Generation:** A structured API query is generated based on user selections.
3. **Response Processing:** The API returns JSON data, which is reformatted into natural language output.

The system ensures that only GBIF metadata is used to generate responses, preventing hallucinated or misleading information. Additionally, rate-limiting strategies were implemented to manage API usage efficiently.

Chat-Based Metadata Exploration

Once the user retrieves datasets matching the query criteria, BotGBIF enables further interaction through a chat-based interface. Users can:

- Request metadata summaries for each dataset.
- Ask specific questions about dataset contents (e.g., "What species are included in this dataset?").
- Retrieve additional contextual information from external biodiversity repositories.

To maintain usability, response length limits were imposed to prevent excessive data overload from large datasets.

Implementation and Testing

The system was iteratively developed and tested using real-world biodiversity queries. The evaluation focused on:

- **Accuracy:** Comparing retrieved results with expected GBIF API responses.
- **Usability:** Assessing ease of interaction and clarity of responses.
- **Performance:** Measuring API response times and LLM processing efficiency.

A pilot test was conducted with biodiversity researchers to refine BotGBIF's capabilities and optimize its response mechanisms. Section 5 presents an application case study demonstrating the system in action.

4 BotGBIF Development

The development of BotGBIF was structured around the goal of bridging the gap between non-technical users and the extensive biodiversity data provided by GBIF. To achieve this, we integrated a LLM capable of interpreting natural language queries and dynamically generating API calls to retrieve relevant information. The development process involved several key phases: API analysis, LLM query structuring, interface design, data processing, and user interaction optimization.

The application is publicly available at https://botgbif.streamlit.app/, and its source code can be accessed in the GitHub repository at https://github.com/gustavomarcelonunez/gbif-streamlit/tree/main.

GBIF API Analysis and Query Structuring

The initial phase involved a detailed exploration of the GBIF API to identify its functionalities and constraints. We conducted systematic testing of various endpoints, analyzed their parameters, and examined the structure of JSON responses. A critical challenge identified during this phase was ensuring that API queries were accurately constructed. While initial attempts allowed the LLM to directly generate API calls, the results were inconsistent due to the model's occasional misinterpretation of required parameters.

To overcome this, we introduced an intermediary layer where user queries were first interpreted by the LLM, then translated into structured API requests through predefined templates. This approach provided greater reliability, ensuring that queries adhered to the API's specifications while still being flexible enough to accommodate diverse user inputs.

Interactive User Interface with Streamlit

For user interaction, we developed a web-based interface using Streamlit, chosen for its ease of implementation and support for interactive applications. The interface allowed users to specify their search parameters manually, including:

– Country (Geographic scope)
– Dataset type (Occurrence, checklist, metadata, etc.)
– Keyword search (Taxonomic group, location, or dataset-specific terms)

Once parameters were selected, BotGBIF retrieved relevant datasets from the GBIF API, displaying concise metadata previews. Unlike traditional search tools that return raw JSON outputs, our system formatted responses into user-friendly summaries, making biodiversity data more accessible to researchers, educators, and conservationists.

Metadata Retrieval and Contextual Interaction

A core feature of BotGBIF is its ability to facilitate interactive exploration of dataset metadata. After retrieving datasets matching the user's criteria, the system enables users to engage in a conversational search experience. Instead of manually parsing complex metadata structures, users can ask questions such as:

– Who contributed to this dataset?
– What is the geographical coverage of these records?
– What species are included in this dataset?

BotGBIF leverages GPT-4 to analyze the retrieved metadata and generate structured responses. To prevent information overload, only the first fifteen datasets matching the search criteria are initially displayed, with the option to refine or expand results based on follow-up queries.

Handling Large Data Volumes and Pagination

One of the challenges encountered during development was the sheer volume of data returned by the GBIF API. Since biodiversity datasets can contain thousands or even millions of records, direct retrieval of full datasets was impractical. To manage this, we implemented:

– Pagination controls, ensuring that only a subset of data is processed per request.
– Metadata prioritization, where only essential fields (e.g., dataset title, contributors, geographic scope) are retrieved initially.
– Dynamic filtering, allowing users to refine results interactively without overloading the system.

These strategies ensured that the application remained responsive while still providing comprehensive access to biodiversity data.

Conversational Query Refinement

A distinguishing feature of BotGBIF is its ability to maintain conversation context and allow iterative refinement of queries. Users can adjust their search dynamically, filtering results further based on taxonomic groups, time periods, or geographic regions. Unlike static query interfaces, this conversational approach enhances the user experience by reducing the need for technical expertise in formulating API requests.

Additionally, since GBIF datasets contain diverse metadata formats, BotGBIF adapts responses based on the dataset structure. If a dataset includes occurrence records, for example, the system can generate visualizations such as georeferenced species distribution tables to aid interpretation.

5 Use Case: Querying Biodiversity Data with BotGBIF

To illustrate the capabilities of BotGBIF, we present a step-by-step use case in which a user retrieves and explores biodiversity datasets using natural language queries. This scenario demonstrates how BotGBIF facilitates seamless interaction with the GBIF database, from dataset discovery to dynamic metadata exploration.

Step 1: Defining the Search Criteria

The user accesses the BotGBIF interface and specifies search parameters through an intuitive interface with dropdown menus and text fields. In this example, the user is interested in biodiversity records from Argentina, particularly occurrence data related to Península Valdés. The selected criteria are:

- Country: Argentina
- Dataset Type: Occurrence
- Search Keyword: "Península Valdés"

Once the parameters are set, the user submits the query, initiating the search process.

Step 2: Retrieving and Displaying Matching Datasets

Upon receiving the query, BotGBIF constructs a request to the GBIF API and retrieves datasets that match the specified criteria. The results are displayed in a structured list, presenting key metadata fields such as:

- Dataset title
- Creation and last modification dates
- Digital Object Identifier (DOI)

The user browses the list and selects a dataset for further exploration (see Figure 2).

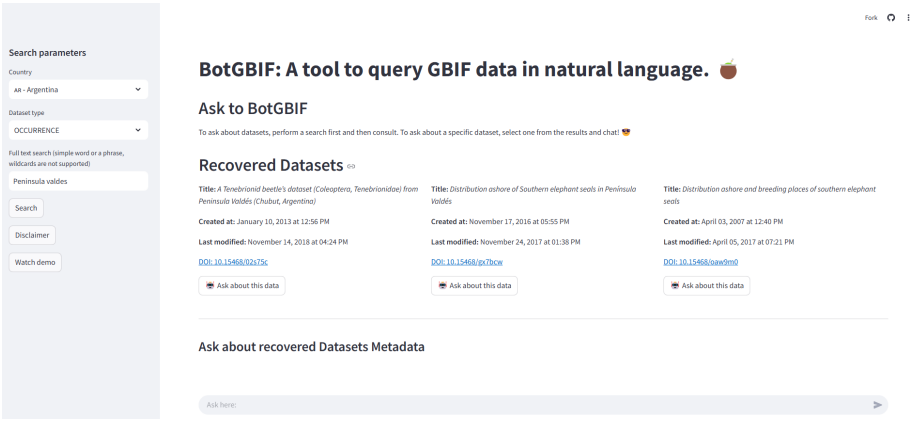

Fig. 2. BotGBIF search interface where users can manually input parameters to retrieve biodiversity datasets.

Step 3: Selecting a Dataset and Initiating Interaction

Among the retrieved datasets, the user identifies an interesting one, for example: Distribution ashore of Southern elephant seals in Península Valdés. By clicking on the dataset, the user initiates a conversational session to explore its metadata interactively.

Step 4: Conversational Interaction with the Dataset

Once the dataset is selected, the user interacts with BotGBIF via a chat interface, submitting queries in natural language to extract specific information. For instance:

– User query: "What species are included in this dataset?"

BotGBIF processes the request, retrieves relevant metadata, and generates a structured response listing the species found in the dataset (see Figure 3).

Step 5: Exploring Additional Insights

The user continues the interaction by asking follow-up questions to extract deeper insights. Examples of queries include:

– How many records exist for each species?
– What is the temporal range of the dataset?
– Are there any associated environmental variables?

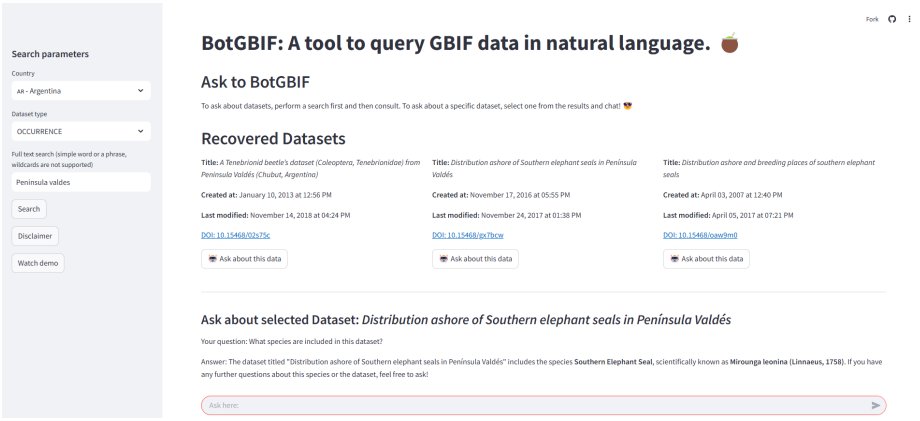

Fig. 3. Chat interface displaying species found in the selected dataset.

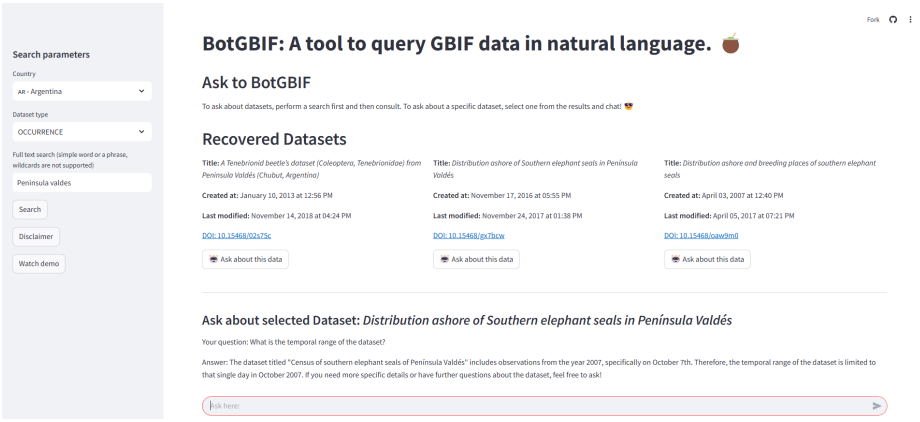

Fig. 4. Chat interface displaying the temporal coverage of the selected dataset.

BotGBIF dynamically processes these requests, retrieves the corresponding data, and presents the results in a clear and structured format. For example, when querying the dataset's temporal coverage, the system provides a summary of the recorded time span, as shown in Figure 4.

This iterative and conversational approach allows users to refine their queries dynamically, making the exploration of biodiversity data more accessible and intuitive. By leveraging LLM capabilities, BotGBIF simplifies metadata retrieval and enables users to gain meaningful insights without requiring technical expertise in API interactions.

6 Discussion

This study explored the integration of LLM with biodiversity databases, particularly the GBIF API, to facilitate natural language interaction. The research was guided by three key questions, which we address in this discussion.

The RQ1 examined whether natural language queries could be effectively answered using only the GBIF API, without relying on external data sources. Our findings indicate that while the GBIF API provides a vast amount of biodiversity data, it is structured primarily for expert users familiar with database queries and metadata interpretation. The reliance on structured endpoints and JSON responses makes direct, unassisted retrieval of information challenging for non-technical users. Through our approach, LLMs improved accessibility by translating natural language questions into structured API requests. However, certain limitations persisted, such as the inability of GBIF to provide contextual explanations or synthesize information beyond its raw data. This suggests that while the GBIF API is a powerful resource, it benefits significantly from an intelligent intermediary capable of restructuring and summarizing responses.

The RQ2 question considered whether an LLM could intelligently utilize additional services to supplement information retrieved from GBIF, ensuring more comprehensive responses. While our implementation focused solely on GBIF, results indicate that incorporating external knowledge bases could improve response completeness and accuracy. For instance, supplementing species occurrence data with ecological traits from databases like Encyclopedia of Life (EOL) or integrating environmental variables from sources such as WorldClim could provide richer insights. Future iterations of BotGBIF could leverage multi-source retrieval mechanisms where the LLM dynamically determines when external data is required, thus overcoming the inherent limitations of a single data provider.

RQ3 explored how LLM integration enhances accessibility and usability for non-technical users, promoting broader adoption and more informed decision-making. A major strength of BotGBIF is its ability to abstract technical complexity, enabling users to interact with biodiversity data conversationally. This lowers the barrier to entry for researchers, educators, and policymakers who may lack programming expertise but require biodiversity insights.

Additionally, the interactive nature of LLMs allows for iterative refinement of queries, ensuring that users obtain the most relevant results. However, challenges remain regarding the accuracy and interpretability of LLM-generated responses, as general-purpose models like GPT-4 are not explicitly trained on biodiversity domain knowledge. Enhancing transparency in AI-generated answers and integrating domain-specific fine-tuning approaches could further improve usability.

7 Conclusions

Accessing biodiversity data is essential for ecological research and conservation but remains challenging for non-technical users due to the complexity of API queries. This study explored the integration of LLMs with the GBIF API to enable natural language interaction, improving accessibility and usability.

Our findings show that LLMs can effectively assist users by transforming natural language inputs into structured API calls, provided that query generation is guided by user-selected parameters. Additionally, LLMs improve data interpretation by summarizing JSON responses in an intuitive format, reducing technical barriers to biodiversity information.

Despite these benefits, limitations remain. The large volume of GBIF data requires constraints on query results, and LLMs may lack domain-specific precision. Scalability and cost are also concerns, particularly for deploying high-performance models like GPT-4.

Future work should focus on training LLMs with biodiversity-specific datasets, improving cost-efficient AI deployment, and integrating additional biodiversity knowledge bases to enhance accuracy. BotGBIF represents a step toward more accessible biodiversity data retrieval, demonstrating the potential of AI in facilitating scientific data access and conservation efforts.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ..., & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cambridge. (2024). Ai-driven conservation copilot: Revolutionising biodiversity solutions. https://ai.conservation.cam.ac.uk/projects/ai-driven-conservation-copilot-revolutionising-biodiversity-solutions/

Chen, L., Wang, J., & Zhang, H. (2022). Advancing scientific data retrieval through natural language processing. *Journal of Scientific Computing*, *45*(3), 112–129.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Domazetoski, V. (2024). *Enhancing ecological knowledge discovery using large language models* [Doctoral dissertation, Master's Thesis, Georg-August-Universität Göttingen].

Eldredge, N. (2000). *Life in the balance: Humanity and the biodiversity crisis*. Princeton University Press.

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures* [Doctoral dissertation, University of California, Irvine].

iDigBio. (2024). Announcing the biodiversity chatbot (new!) https://www.idigbio.org/home

Ladle, R. J., & Whittaker, R. J. (2020). Biodiversity and conservation. *Annual Review of Ecology, Evolution, and Systematics*, *51*, 233–258.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6), bbac409.

Nejjar, M., Zacharias, L., Stiehle, F., & Weber, I. (2025). Llms for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process*, *37*(1), e2723.

Norway, G. (2024). Chatipt. https://chatipt.svc.gbif.no/

OpenAI. (2025). Chatgpt (versión gpt-4) [Accedido el 15 de abril de 2025].

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L., & Desmet, P. (2014). The gbif integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PloS one*, *9*(8), e102623.

Touvron, H., Hwang, J. L., & Misra, I. (2023). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Vert, J.-P. (2023). How will generative ai disrupt data science in drug discovery? *Nature Biotechnology*, *41*(6), 750–751.

Von Wettberg, E., & Khoury, C. K. (2022). Biodiversity data: The importance of access and the challenges regarding benefit sharing.

Wang, S., Liu, C., & Patel, R. Ai-mediated data retrieval: Approaches and challenges. In: In *International conference on artificial intelligence and scientific computing*. 2023, 201–215.