DCMS AI Pilot Project Report

Herbarium in Dialogue: An LLM-Powered Chatbot of Professor Samuel Pickworth Woodward

Natural History Museum, London

Naifeng Zhang[1], Elise Gallois, Sanson T. S. Poon, Arianna Salili-James, and Ben Scott

Royal Agricultural University

Kelly Hemmings, Cassie Newland

31st March 2025

1 Project

1.1 Aim

The project aims to create a chatbot utilising Large Language Models [LLMs, 8] to provide in-depth, contextually rich responses centred on the historic herbarium of Prof. Samuel Pickworth Woodward. This expert model will adopt the persona of Prof. Woodward, who was a lecturer in Geology and Natural History at the Royal Agricultural College (RAC) from 1845 to 1848 [9]. The chatbot is designed to engage users with informative and historically insightful content, leveraging botanical data from Prof. Woodward's collection to facilitate interactive inquiries related to botanical interests. The primary goal is to tailor the LLM using techniques such as prompt engineering [5] and Retrieval-Augmented Generation [RAG 2] to ensure accurate, interesting and educational responses. This initiative seeks to broaden public engagement with the natural sciences by providing a reliable and ethically sound platform. Key features include the use of verified scientific names for species identification, the application of place names and decimal coordinates for location-based queries, and the integration of both image-based and textual information regarding herbarium specimens. By showcasing Prof. Woodward's significant contributions and his extensive collection of early specimens from the UK and abroad, the project aims to enhance user interaction and foster a deeper appreciation for this invaluable botanical collection.

[1]naifeng.zhang@nhm.ac.uk

1.2 Background

In 2023, the Royal Agricultural University (RAU) initiated the digitisation of its herbarium, capturing both text and images. This project was made possible through funding from the Gloucestershire Naturalists' Society, The Cirencester Fund, and the RAU Innovation Fund. The primary objective was to publish the resulting metadata and images on platforms such as the National Biodiversity Network and the Global Biodiversity Information Facility, thereby expanding access to these valuable botanical resources.

Digitised herbaria have been shown to increase the reach and accessibility of natural science collections [1], but their engagement has largely remained within traditional academic and specialist circles, such as researchers and students in botany. To explore alternative ways of engaging with a broader audience, we proposed the development of an Artifical Intelligence (AI)-driven chatbot that could leverage the digitised herbarium's metadata and images. While a few natural science chatbots exist, such as DeepAI's notable Darwin Chatbot, we did not find any herbarium-specific examples. This project aimed to fill that gap by trialling a chatbot that interacts with users using an authentic, curated dataset rather than relying solely on generic publicly-available information. Unlike chatbots designed for rigorous academic research, our focus was on fostering general interest and public engagement with botanical heritage.

The Woodward collection provided an ideal test case due to its manageable size (approximately 5,000 specimens), taxonomic diversity (around 250 families and 1,600 species), personal connection to the RAU, and extensive network of contributing collectors (around 300 individuals). Predominantly composed of UK-collected wild specimens, with a small proportion of international samples, the collection also stood out for the age and visual appeal of its specimens—factors that could help attract new audiences.

1.3 Method Outline

To power the chatbot, we turned to Large Language Models [LLMs 8]. LLMs, such as ChatGPT [Chat Generative Pre-trained Transformer 3], have revolutionised AI by enabling machines to generate human-like text based on patterns learned from vast datasets. While LLMs are highly adaptable, their general-purpose nature often limits their effectiveness in specialised domains. To ensure that the chatbot provided accurate, relevant, and contextually appropriate responses, LLM customisation was necessary.

Customising LLMs for domain-specific applications involves several techniques, including prompt engineering, Function calling, and Retrieval-Augmented Generation (RAG). Prompt engineering refines the input prompts given to the model, guiding responses to align with the desired context and format [5]. Function calling allows the model to interact with external databases or APIs, enabling dynamic data retrieval and real-time updates [4]. RAG further enhances the chatbot's accuracy by supplementing the model's built-in knowledge with verified, domain-specific datasets, ensuring responses remain grounded

in factual information [2]. These techniques are crucial for applications of historical and scientific exploration, where accuracy and credibility are paramount.

2 Dataset

2.1 Raw dataset

A major part of this project is the metadata behind the Samual Pickworth Woodward herbaria based at Royal Agricultural University. This data follows Darwin Core standards [10], and involves the manual transcription of herbarium sheet labels, capturing species names, collection locations, dates, and associated botanists. Taxonomic family names were also transcribed from the original herbarium folders that housed the genera within each family. A numbering system was implemented to identify each folder and sheet, accommodating historical practices where multiple specimens were mounted on a single sheet. Expert botanists, including Botanical Society of Britain and Ireland determiners, county recorders, and field key authors, verified species identities and documented any taxonomic reclassifications, achieving 85% verification. Specimens were georeferenced using ISO two-digit country codes, vice-county codes for GB and Ireland, and decimal latitude and longitude with an uncertainty radius in metres. Throughout the process, data accuracy was continuously checked alongside species verification.

To enhance the dataset's accessibility for this project, additional steps were taken, including adding a data field for species common names and standardising the names of collectors, contributors, and exchangers, as the original metadata contained variations such as with initials, full names, and titles. A second round of data verification was conducted simultaneously to ensure consistency and accuracy. The final raw dataset comprises 5,030 rows, each representing a unique specimen and including information on taxonomic identification, collection details, and georeferenced data.

The dataset also include 2,776 high-resolution herbarium sheet images. These photographs were captured using a Canon M50 Mark II camera, with the setup involving an LED light box, a camera stand, and a remote shutter. As illustrated in Figure 1, each image features a Golden Thread scale rule for colour calibration, and the RAU logo with a copyright notice. The RAW images were later converted to JPEG format using Canon Photo Professional software. Each image corresponds to one or more metadata entries.

2.2 Data Preprocessing

Pre-processing was done on the original dataset to ensure that the data would be as legible as possible for the LLM. For example, in order to avoid misinterpretations in regards to dates, the metadata was formatted to split day, month, year into separate columns. Blank spaces and "N/A" values were also used to handle missing or uncertain data, for additional clarity.

Figure 1: An example herbarium sheet image from the RAU collections, featuring *Luronium natans* L. (Alismataceae) collected by J. Price in 1848 at Ellesmere Lake, Shropshire, UK. Identified by Clare and Mark Kitchen in 2024. The image includes a Golden Thread scale rule (right edge, for colour calibration) alongside the RAU logo with copyright notice.

3 Methods

The chatbot that forms this project is based on a customisation of a pre-trained large language model. In our project, we employ a combination of Retrieval-Augmented Generation (RAG) and prompt engineering to enhance the model's accuracy, relevance, and contextual depth. This approach ensures that the model can dynamically retrieving and integrating relevant external knowledge (herbarium sheet metadata and images) as well as providing domain-focused (herbarium and botany related) responses. The source code for this work has been made available on Github[2].

3.1 Function-Enhanced Retrieval-Augmented Generation

RAG addresses the limitations of static internal knowledge in LLMs by dynamically retrieving external information from structured data and documentation. The process involves three key steps:

1. Document embedding and vectorization. This involves converting knowledge sources into dense vector representations.
2. Semantic search and retrieval. The matching of user queries against indexed sources to fetch relevant documents.
3. Response generation. In other words, where retrieved content is fused with the LLM's output to enhance factual accuracy and depth.

However, simple vector-based semantic search may not meet our project's needs, par-

[2]github.com/NaturalHistoryMuseum/nhm_dcms_rau_herb

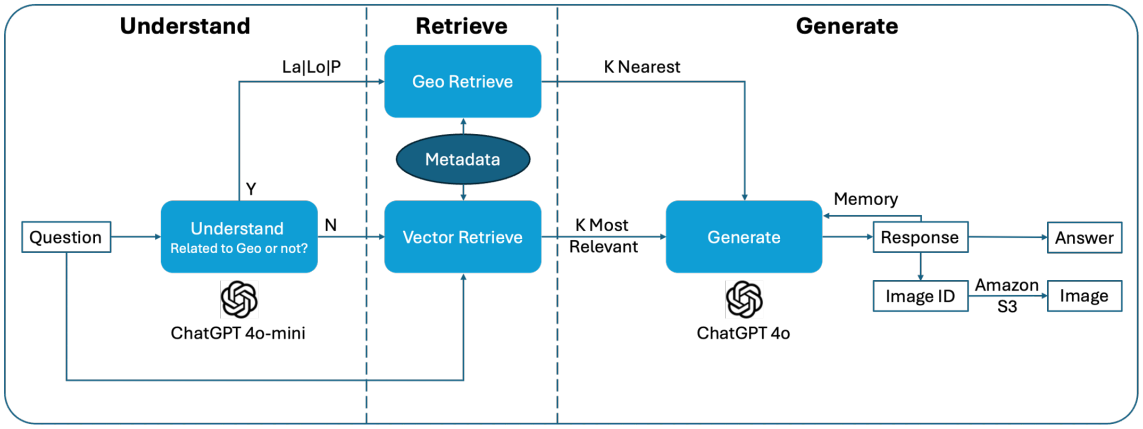

Figure 2: Enhanced RAG system with Function Calling integration. The workflow is comprised of three stages: (1) **Understand**: A lightweight LLM (ChatGPT 4o-mini) detects geographical intent; (2) **Retrieve**: If applicable, a function retrieves the nearest K metadata entries; otherwise, a vector-based search fetches relevant data; (3) **Generate**: ChatGPT 4o generates responses, including an Image ID for retrieving related images via Amazon S3.

ticularly for georeferenced data. To address this, we integrated Function Calling—a capability that allows LLMs to dynamically invoke pre-defined tools by generating structured requests—into our enhanced RAG system, structured into three distinct stages for precise, context-aware responses:

Understand: Context Analysis A lightweight LLM [ChatGPT 4o-mini 7] first analyses the user's query to determine its geographical relevance. If a location is identified (e.g., *'Tell me about a specimen collected in Cirencester'*), the model generates structured location information in the format `La|Lo|P`, where `La` and `Lo` represent decimal latitude and longitude, and `P` is the place name. Otherwise, the system defaults to standard vector-based retrieval.

Retrieve: Context-Adaptive Data Retrieval For place-related queries, a function retrieves the K nearest metadata entries using geospatial distance calculations. This function computes the distances between the query coordinates and all metadata locations, ultimately identifying the K nearest specimens. For non-geographical queries, the system employs a vector-based semantic search. Both the query and the metadata are transformed into embeddings using OpenAI's text embedding model. The cosine similarity between each metadata entry and the query—representing their contextual similarity—is calculated, and the K most relevant metadata entries are retrieved. Additionally, retrieval is enhanced by incorporating time-awareness, as illustrated in Figure 3. This refinement improves responses to temporal queries, such as *'What is the oldest specimen in your herbarium collection?'* or *'Did you collect any specimens after 1900?'*.

Generate: Multimodal Response Finally, we use ChatGPT-4o [6] to generate comprehensive responses that include both textual answers and the 'Image ID' of the referenced metadata. This ID enables retrieval of corresponding images stored in Amazon

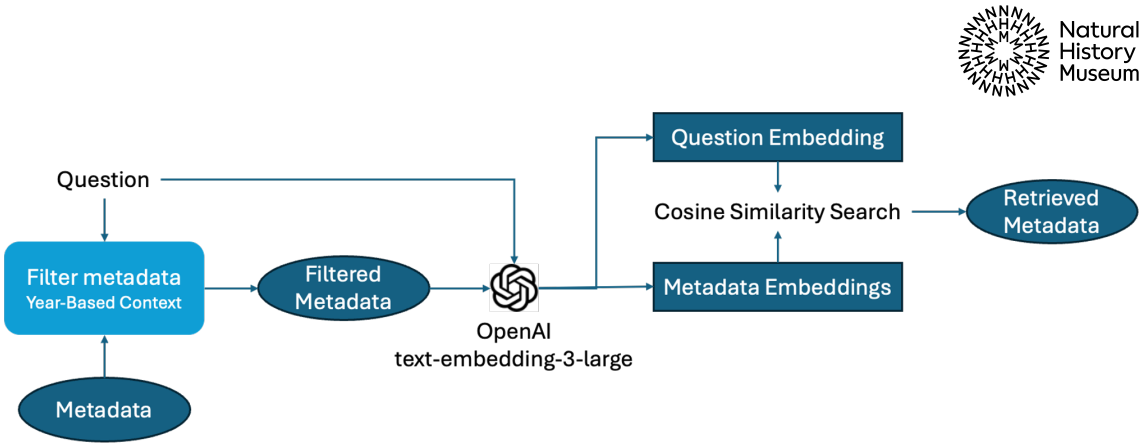

Figure 3: Similarity search with year-based context. In the 'Vector Retrieve' component (Figure 2), the metadata is first filtered by the user's temporal constraints. The query is then embedded and compared via cosine similarity against filtered metadata embeddings to retrieve the most relevant results.

S3 cloud storage, adding a visual dimension to the response.

By integrating a function calling approach, our system ensures a seamless transition between geospatial and non-geospatial workflows, dynamically adapting retrieval strategies based on the context of user queries. This enhancement not only improves data retrieval accuracy but also enables a more flexible and context-aware response generation. However, to fully leverage these capabilities, careful prompt engineering is essential. By crafting precise and structured prompts, we can effectively guide the model to appropriately integrate the retrieved metadata, hence obtaining high-quality responses. The next section delves into the strategies and techniques during the prompt engineering work.

3.2 Prompt Engineering

Prompt engineering involves designing and optimising prompts to guide the LLM towards generating more precise and contextually appropriate responses. Common methods include:

- **Few-shot and zero-shot prompting:** Utilising example-based or direct instruction-based prompts to improve model performance.
- **Structured prompt templates:** Implementing predefined templates to standardise responses for consistency or to facilitate integration with subsequent components in the RAG system.
- **Context injection:** Incorporating domain-specific information into the prompt to ensure the model remains within the expert domain.

Prompt engineering is a multi-round iterative process involving refinement based on the model's output. Each round of refinement addresses issues identified in the model's responses or incorporates feedback from herbarium experts at RAU. The final prompt is the result of six months of iterative development and refinement. We will now detail the key components of the final prompt that achieve the desired functionality and tone.

Data Injection: The conversation history, location information and retrieved metadata are injected and integrated with the prompt. **Persona:** The prompt instructs the

LLM to embody the persona of Samuel Pickworth Woodward, characterized by the curiosity and enthusiasm of a Victorian scientist. It also requires a blend of herbarium knowledge with an engaging tone typical of the era. **Wording Style:** The language is crafted to combine scholarly insight with accessibility, essential for audience engagement while reflecting the persona's passion for botanical sciences. This involves the use of historical phrasing and terminology that resonates with 19th-century dialogue, yet remains comprehensible to contemporary listeners. **Location Handling:** When the user provides detailed information about a specific location, the prompt instructs the model to utilise the calculated distances from the metadata and naturally integrate this data into the response. If the user does not specify a particular location, the prompt requires the model to first inquire with the user, demonstrating an interactive and user-oriented approach. **Bias Control:** Instructions explicitly guide the avoidance of gender-specific language, insisting on gender-neutral terminology to mitigate biases. **Topic Steering:** Responses are centred around botany, with prompts designed to naturally guide off-topic conversations back to the main subject. **Examples for Structured Output:** The prompt outlines specific formatting rules, such as inserting Image ID metadata at the beginning of the response, including determiner information at the end, and using italics for plant names. Few-shot prompting is employed here, with examples provided to enhance the LLM's understanding of the required format and ensure structured responses that will be utilised in our RAG system.

4 Model Evaluation

Evaluating large language models presents unique challenges, as they lack standardised performance metrics seen in traditional machine learning tasks. Given the objectives of our project, which involve assessing chatbot performance across multiple dimensions, we designed a structured evaluation framework consisting of 30 carefully curated questions. Notably, these questions were initially generated by ChatGPT itself, with subsequent refinements made by collaborators at RAU, to enhance clarity and relevance.

The evaluation questions are spit into six categories:

- Metadata Integration (10 questions) – Assesses how well the model integrates metadata, provides historical insights, and responds appropriately.
- Georeferenced Data Handling (4 questions) – Assesses if the Function-Enhanced RAG system can correctly understand the geospatial context of the question, retrieve right metadata, and integrate the data in its responses seamlessly.
- Off-Topic Handling (4 questions) – Tests how well the model manages irrelevant or out-of-scope queries.
- Bias Control (4 questions) – Measures the model's ability to mitigate and avoid biases in responses.
- Misdirection & Trap Questions (4 questions) – Determines the model's robustness

against misleading or adversarial inputs.

- Ethical Considerations (4 questions) – Evaluates the model's responses in ethically sensitive contexts.

To better demonstrate the chatbot's capabilities with the improved prompt and RAG system, we performed comparative experiments on models with varying degrees of degradation. This involved testing five different models, progressively reducing prompt complexity and external information retrieval:

- Model 1 (Full Prompt): The model uses the complete refined prompt along with metadata from the function-enhanced RAG system.
- Model 2: This stage uses a similar prompt to Stage 1 but removes any geospatial components from both the prompt and the RAG system.
- Model 3: The prompt remains the same as in Stage 2, but the RAG system no longer considers year-based context. Instead, it directly employs vector-based similarity search.
- Model 4: Metadata is removed, while instructions on styling, persona, bias control, and ethical considerations are retained.
- Model 5 (Minimal Prompt): The model functions in its simplest form, relying only on ChatGPT with a length constraint.

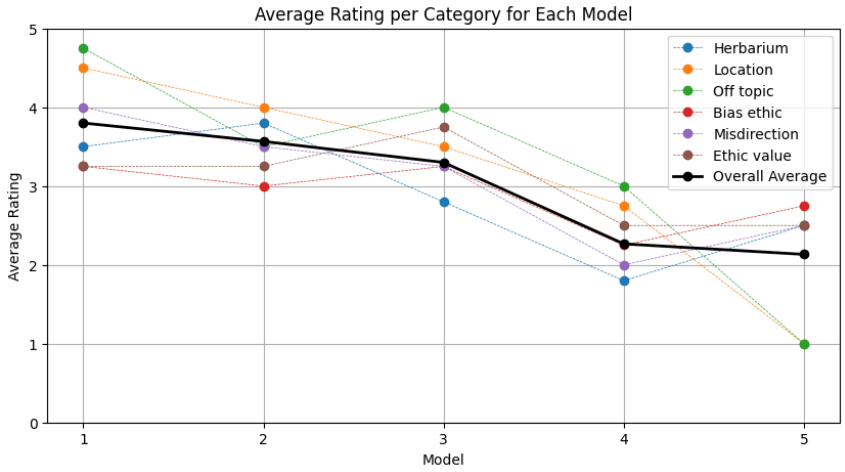

Figure 4: The average ratings of five models across six evaluation categories (Metadata Integration, Georeferenced Data Handling, Off-Topic Handling, Bias Control, Misdirection & Trap, and Ethical Considerations). Colour-coded dashed lines show category trends, while the black solid line marks the overall average scores of each model.

We asked herbarium expert Kelly Hemmings from RAU to evaluate the responses of each model across all 30 questions, assigning scores from 1 (worst) to 5 (best). The overall average scores and category-wise averages for the five models are presented in Figure 4. A clear downward trend in the overall average scores (solid black line) indicates that our model generally outperformed the others on these questions. To further analyse performance across question categories, we provide detailed scores for each model in Figure

5.

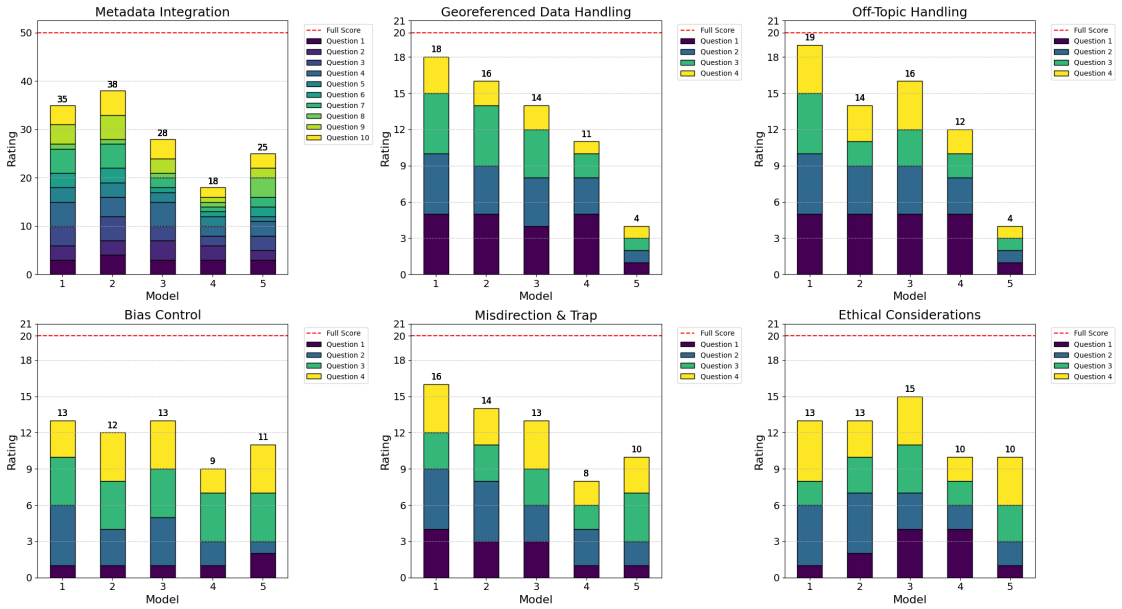

Figure 5: Comparative visualisation of five models across six categories of questions. Bar charts show question-specific ratings (colour-coded) and total category scores displayed above. The red dashed line represents the maximum possible score for each category.

In evaluating metadata integration, Model 4 underperformed compared to Model 5 due to its restriction in prompt that limit its historical knowledge up to Prof. Woodward's era, lacking awareness of modern advancements. Model 3 demonstrated a notable improvement over Model 4 by incorporating contextual metadata, which enhanced its responsiveness. Models 1 and 2 surpassed Model 3 overall, particularly excelling in handling year-specific queries linked to herbarium records, thereby elevating their performance in herbarium related inquiries.

Regarding location-based queries, Model 4 outperformed Model 5 as it could address basic location-based questions using inherent knowledge, though it lacked nuanced details such as collector names or precise dates. Models 2 and 3, enriched with metadata containing exact geographical descriptors, delivered stronger responses by leveraging place-name references through vector-based search methods. Model 1 achieved the highest accuracy, adeptly resolving broader spatial queries—for example, identifying specimens from "the southernmost UK"—by utilising location-based metadata retrieving functionality.

On bias control, models incorporating prompt restrictions and metadata generally managed bias-related risks effectively. However, an overly rigid directive for gender-neutral language in Model 1 led to irrelevant responses when asked about women's historical contributions to herbarium collections, resulting in a low score. Refining the prompt to "avoid stereotypes or assumptions favouring any gender" may yield more balanced, context-aware answers.

For off-topic queries, Model 1 excelled by politely acknowledging digressions before

redirecting focus to herbarium-related details. Its concise, structured responses—free of overly elaborate language—contrasted with Models 2 and 3, which occasionally prioritised stylistic flourishes over substantive content or provided insufficient botanical context.

In addressing misdirection&trap questions, Model 1 remained preferred for its tactful navigation of irrelevant prompts, steering conversations back to botany while offering specific herbarium insights. Model 2, though coherent, was marked down for excessive verbosity, which diluted clarity and precision.

Finally, on ethical considerations, Models 1–3 outperformed Models 4 and 5, underscoring the value of integrated metadata in addressing issues such as colonial-era collection practices. Access to contextual records enabled nuanced discussions of historical ethics, whereas Models 4 and 5, reliant solely on general knowledge, produced vaguer responses.

In summary, our model generally outperformed the others across all evaluation categories, demonstrating superior accuracy in metadata integration, geospatial queries, and ethical considerations. Its ability to leverage contextual metadata allowed for precise responses, particularly in year- and location-based inquiries. While its strict bias-control measures were generally effective, overly rigid gender-neutral directives occasionally led to irrelevant answers. Our model also excelled in handling off-topic and misdirection questions by maintaining botanical relevance and providing concise yet informative responses. Its strong performance underscores the value of integrating structured metadata and refined prompt engineering for enhancing chatbot reliability and contextual awareness.

5 Discussion and next project phase

5.1 Summary

During this project, we successfully developed an AI-based chatbot depicting Prof. Samuel Pickworth Woodward. Employing advanced techniques such as Retrieval-Augmented Generation, function calling, and prompt engineering, the chatbot provides accurate and contextually detailed answers about herbarium specimens. Additionally, the model's inclusion of georeferencing, and specimen images, helps it stand out amongst other notable persona chatbots, and enhances the interactive user experience. Furthermore, the feedback and evaluation from the experts at RAU suggests that the chatbot effectively generates accurate responses based on herbarium data whilst also fostering public interest in herbarium and botany in an engaging manner.

5.2 Challenges & Next Steps

Despite its success, during this project, we faced certain challenges limitations with the chatbot. For example, the dependency on the genearation of decimal latitude and longitude for specific locations by the LLM could lead to inaccurate georeferencing, that may be resolved in future by incorporating geolocation tools such as Google Earth. Furthermore,

the existing retrieval method sometimes struggles with complex temporal or quantitative queries. This is something that can be improved by incorporating SQL queries into the RAG system. Additionally, the system is not yet optimised to manage broad and intricate queries such as expansive questions on climate change, that may be addressed by further fine-tuning in collaboration with experts in the field.

5.3 Impact & Next Steps

LLM-based chatbots can no doubt play great roles in many sectors including the natural sciences. This is emphasised by the recent developement of the Biodiversity Chatbot that utilizes iDigBio and GBIF collections. The goals of our chatbot here are within herbarium education and outreach. The final phase of this project involves the deployment the chatbot within Royal Agricultural University, not only to make it publicly accessible online but to also utilize it within teaching and outreach at RAU. Additionally, the chatbot will also play a part in the important 180 Year Anniversary celebrations of the founding of RAU in 2025. Furthermore, there is potential to expand the project by inviting participation from other academics and research institutions with herbarium collections, in order to build a more comprehensive chatbot. By incorporating herbarium data from multiple sources, the chatbot can be enhanced to provide richer, more diverse responses, increasing its educational value and scope. Beyond herbaria, there's potential to expand the use of this chatbot technology to other areas, such as museum exhibits, educational platforms, and other scientific domains. This expansion could involve collaboration with institutions such as Natural History Museum, to adapt the chatbot for different content while building on the foundational technology developed in this project.

References

[1] H. Hardy, L. Livermore, P. Kersey, K. Norris, and V. Smith. Understanding the users and uses of uk natural history collections. *Research Ideas and Outcomes*, 9:e113378, 2023.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[3] OpenAI. Chatgpt: Chat generative pre-trained transformer, 2023. URL `https://chatgpt.com`. Large language model.

[4] OpenAI. Function calling: enable models to fetch data and take actions, 2023. URL `https://platform.openai.com/docs/guides/function-calling`.

[5] OpenAI. Prompt engineering: enhance results with prompt engineering strategies, 2023. URL https://platform.openai.com/docs/guides/prompt-engineering.

[6] OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

[7] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

[8] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[9] W. A. S. Sarjeant. *Geologists and the History of Geology: An International Bibliography from the Origins to 1978*. Arno Press, New York, 1st edition, 1980.

[10] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. Darwin core: An evolving community-developed biodiversity data standard. *PLOS ONE*, 7(1):1–8, 01 2012. doi: 10.1371/journal.pone.0029715. URL https://doi.org/10.1371/journal.pone.0029715.