Leveraging large language models for ecological interpretation using an eBird chatbot case study

Elise C. Gallois [1,2], *Arianna Salili-James* [1], *Sanson T. S. Poon* [1], *Artur Trebski* [1], *David W. Redding* [1]

1) *Science Department, Natural History Museum, London SW7 5BD, UK*
2) *UK Centre for Ecology and Hydrology, Edinburgh Research Station, Penicuik, UK*

Corresponding author email: eligal@ceh.ac.uk

Abstract

1. The anthropocene presents significant challenges for global biodiversity, public health, and long-term ecosystem stability. The wealth of publicly available near-real-time ecology and climate data can be used to monitor these challenges and allow practitioners to develop mitigation strategies.
2. There is untapped potential to apply Large Language Models (LLMs) to quantitative ecological and environmental datasets, enabling researchers and practitioners to use natural language queries to transform ecological observations into actionable insights for both conservation action and external communication of results to diverse audiences. Advances in artificial intelligence (AI), and particularly in LLMS, offer emerging opportunities to address these challenges. LLMs are increasingly proficient at identifying patterns and semantic relationships within textual data, and are highly customisable. Accessible AI tools can also facilitate communication across research and policy sectors.
3. Here, we present a roadmap for designing and implementing multi-modal LLMs to answer ecological research questions. In order to build 'virtual statistician' systems capable of fast-tracking data interpretation, we advocate for strategic planning, data stewardship practices, careful prompt-engineering, and model evaluation as key steps in the LLM development process.
4. We showcase a case study that applies the open-source LangChain framework to analyse citizen science data using the eBird database to produce a chatbot allowing the user to ask quantitative questions about near-real-time bird observations. Using our LLM roadmap, we highlight the importance of iterative and strategic prompt engineering and agent selection, in addition to iteratively evaluating model output.
5. As LLM software continues to evolve, their integration into ecological and environmental research can empower ecologists with purpose-built tools that bridge the gap between data collection and actionable solutions.

Keywords: *large language models, citizen science, artificial intelligence, natural language processing, multi-agent models*

Introduction

The anthropocene continues to offer novel and unprecedented challenges for global biodiversity, public health, and ecosystem stability (Bellard et al., 2012; Doney et al., 2012; Willis & Bhagwat, 2009). While the size and hierarchical complexity of ecological and social data have increased at a rapid rate, tools to investigate and communicate emerging phenomena within these datasets remain time-consuming and specialised. Artificial intelligence (AI) data tools could facilitate both a democratisation of data analysis and the step-change in pace required to identify emerging trends and prompt rapid intervention responses. Large Language Models (LLMs) are complex probabilistic generative AI Natural Language Processing (NLP) models adept at recognising meaning and identifying semantic interconnectedness and patterns within text. LLMs such as Google's BERT (Bidirectional Encoder Representations from Transformers) and OpenAI's GPT (Generative Pretrained Transformer), and DeepSeek, have been evolving exponentially over the last decade (Google, 2024; OpenAI, 2024; Topsakal & Akinci, 2023; Liu et al., 2024). While this expansive evolution may pose challenges for long-term reproducibility, it also presents lucrative opportunities for scientific efficiency. For instance, LLMs have been used to auto-generate patient discharge forms based on incredibly basic prompts provided by humans (Chatterjee et al., 2023), extract data from survey responses from patients (Haag et al., 2023), and answer complex questions about human genomics to a degree of professional accuracy (Jin et al., 2024). Within the field of ecology, LLMs have been employed to scout bodies of academic text to recognise and report meaningful occurrences of taxa names (Le Guillarme & Thuiller, 2022), identify occurrences of pest control activity (Scheepens et al., 2024), perform biodiversity literature searches using keywords (Abdelmageed et al., 2023) and extract key metadata about pathogen hosts (Gougherty & Clipp, 2024).

There is considerable untapped potential to use natural language processing on structured environmental and ecological quantitative datasets (or, *matrix data* such as CVS, XLS files), for example through the use of open-source software libraries such as LangChain which allow a chat-based interface between existing LLMs and data (Topsakal & Akinci, 2023), or foundational transformer models such as TabPFN which are trained directly on tabular data (Hollman et al., 2025). LLMs as an academic research tool have seldom been applied to quantitative data. Using both historical and 'near-real time' ecological and environmental data as a textual context for AI could offer researchers the opportunity to turn real-time ecological observations into meaningful academic and policy deliverables (Pollock et al., 2025). For example, citizen science data could be an ideal source for harnessing the potential of LLMs (Enríquez-de-Salamanca, 2025). Large open-access global datasets such as iNaturalist (2024) and eBird

(Sullivan et al., 2014), constantly updated by citizen scientists and moderated by subject experts, are already essential tools for researchers studying global biodiversity change, phenology, and species invasion (Chandler et al., 2017; iNaturalist, 2024; Sullivan et al., 2014). By interpreting large amounts of publically available quantitative ecological data, LLMs could enable us to effectively communicate with our datasets, fast-track data interpretation, and lead to actionable conservation and research outcomes (Ceccaroni et al., 2019, 2023; McClure et al., 2020; Pollock et al., 2025). By combining LLMs with existing and robust statistical frameworks and using bespoke NLP tools, it may be possible to create custom multi-modal AI systems which can draw from multiple data sources, which in turn can help ecologists inform conservation decisions and fast-track communication between researchers and policy-makers.

In this paper, we present a roadmap for developing custom multi-modal LLMs to serve as virtual data assistants—or 'virtual statisticians'—designed to support ecologists in summarising, visualising, and exploring trends within complex ecological datasets. These tools represent a timely and powerful opportunity for ecological researchers to interact with data in more intuitive and accessible ways. In the Methods section, we outline a novel and flexible protocol for integrating ecological and environmental matrix data into tailored LLM systems. We showcase a case study that applies this protocol to develop and iteratively refine a LangChain-powered AI model. This model functions as an interactive chatbot trained on the eBird citizen science database, allowing users to ask natural language questions about near-real-time bird observations—including species-specific trends and spatial distributions. In this section, we explore the following research questions:

1) Can a chatbot app using LangChain and a pre-trained OpenAI LLM allow us to interact with citizen science matrix data in a scientifically meaningful way?
2) How well does the model perform using different types of ecological query topics?

Finally, we explore how multi-modal LLMs could be used more broadly across ecological research and conservation practice. We argue that now is a critical moment for ecologists to shape and adopt these tools to bridge the gap between large, complex datasets and timely, actionable insight.

Methods

1. Designing robust and effective quantitative LLMs

This section outlines a structured approach to develop an application to integrate data processing, AI models, and user interaction. The entire approach is represented as a visual roadmap in **Fig.1** and is then used to showcase the design, implementation, and evaluation of a working citizen science chatbot

(Methods Section 2). In **Phase 1** of our roadmap, we gather and preprocess relevant data, selecting appropriate sources and addressing any potential biases or gaps. **Phase 2** involves designing and refining AI agents through prompt engineering, followed by iterative testing to ensure accurate and effective responses. In **Phase 3**, we focus on integrating and deploying the system, ensuring it performs reliably in real-world scenarios. Throughout each phase, continuous evaluation and refinement are conducted to optimise performance and ensure the system's overall effectiveness.

Creating retrieval augmented generation models in LangChain

As LLMs become increasingly integrated into various academic and commercial applications, there is a growing need for frameworks that allow developers to connect these models with bespoke data sources and to create interactive systems. Retrieval augmented generation (RAG) models combine pre-trained generative AI models with the retrieval of selected documents, such as PDF files, text from web-searches, and numerical matrix data (Jeong, 2024; Lewis et al., 2020). LangChain is an open-source RAG software framework designed to enable users to integrate existing pre-trained LLMs (such as OpenAI's GPT models) with a variety of data sources, including matrix data which can be stored as a CSV or SQL dataframe (LangChain, 2024). The LangChain framework also includes the LangSmith developer platform which allows developers to trace runtimes of their models, and LangGraph, an orchestration framework that allows developers to build more complex agentic systems with self-reflective capabilities (LangGraph, 2024). The foundation of LangChain is built on 'chains', which function as chronological query-to-output pipelines. A user provides an informative prompt, along with data inputs (e.g., dataframes, PDFs, or text scraped from web searches), memory inputs from previous model calls, the LLM, and any additional custom tools. Non-academic use cases of LangChain include the development of AI-driven spreadsheets that optimise pricing and automating real-estate operations workflows (LangChain, 2024), and designing intelligent urban traffic control tools (Chen & Ding, 2025).

Prompt engineering and model parameterisation

A 'prompt' is an input that is supplied to an LLM and includes the query from the user in addition to additional instructions provided by the developer, and can therefore be understood as a 'mission statement' for your RAG model. Prompts can also be adjusted to include specific instructions for the LLM, such as scraping the provided text for particular keywords, or to pay particular attention to certain aspects of the data (Scheepens et al., 2024). 'Chain-of-thought' or 'least-to-most' prompting strategies can provide a framework to decompose a user query into a list of easier sub-questions which can be sequentially resolved until the model generates its final output (Zhou et al., 2023). Furthermore, example question and answer sets can be provided within the prompt to guide the model toward an appropriate response (Topsakal & Akinci, 2023). When using quantitative matrix data, the metadata descriptions of the field can be provided in full as part of the prompt to ensure the model is correctly selecting the appropriate variables for analysis based upon the user query. Within LangChain, developers can efficiently build prompts and attach them to their base models using ‘Prompt Templates’ whereby the prompt instructions are included as a text string (LangChain, 2024; Topsakal & Akinci, 2023). Prompts can be iteratively adapted during the development and evaluation stages (Ambrogi, 2023; **Fig.1**), and are integral components to bespoke and advanced RAG models.

Custom LLM tools for data summarisation and visualisation

One of the key advantages of LangChain is the ability to design and attach custom tools and agents to an LLM application. Tools are Python functions that perform a distinct action (such as plotting quantitative data) and are executed when selected by an LLM ‘agent’ which acts as a decision-making component that reads the user input and pre-designed prompt and routes the query to the appropriate tool (Jeong, 2024; LangChain, 2024; Topsakal & Akinci, 2023). A suite of toolkits and agents exist that can enhance the performance of LLM apps designed to process quantitative data, including the CSV and SQL toolkits that optimise agent interactions with quantitative data and execute mathematical queries using Python or SQL code (LangChain, 2024). The GitHub toolkit can connect an LLM app to a provided repository and interact with code, data, and issue tabs. The WolframAlpha tool can connect LLM chains to the WolframAlpha computational search engine to facilitate the computation of more complex mathematical tasks. We recommend using agents and tools for summarising, visualising, and performing mathematical operations on ecological and environmental matrix data within RAG LLM models. Combined with clear and informative prompts, agentic models can receive a user query, design a workflow, assign quantitative functions to either existing or custom-made toolkits, and generate output that is informed by existing metadata.

Orchestration of multiple tools and text sources in LangGraph

LangGraph is a module released by LangChain that allows developers to customise their LLM apps further using an orchestrated and cyclic framework of agents (Jeong, 2024; LangGraph, 2024). Different agents can interact through unconditional (direct, non-optional) or conditional (optional, router-driven) nodes, with memory from the previous agent carried across to the next until a reasonable query has been generated and presented to the user. For quantitative researchers, one key benefit of this system is the ability to draw upon multiple data sources within one app. For example, the developer can build a ‘query routing strategy’ tool that interprets the initial user query and directs it to either an SQL or CSV agent connected to quantitative data, a standard NLP agent drawing upon bank of academic literature stored as PDFs, or even direct it to a web-scraping search tool such as Tavily (Ambrogi, 2023; Gao et al., 2024; Jeong, 2024; LangGraph, 2024). Through prompt engineering and the use of API pulls and real-time web searches, this system provides ecologists and environmental scientists with the opportunity to design, evaluate, and deploy advanced LLM models with conditional logic flows that could help answer user queries about complex ecological phenomena.

Roadmap to effective LLM app development and implementation

There are currently no guidelines for the development and evaluation of LLM RAG models for quantitative researchers. Here, we present a roadmap, split into three phases, for the development of such an app from start to finish (**Fig.1**).

Phase 1: Data gathering and strategic planning

a) <u>*Data Gathering*</u>: Select the quantitative data-frame you would like to provide as the key data source for your app. If available, collate all of the metadata explaining data provenance (e.g. eBird citation), variable names (e.g. observation counts) and units (e.g. metres). You may choose a static data-frame to upload manually to your coding environment. You may alternatively choose to call 'near-real-time' data from an API (e.g. iNaturalist API or the Global Health Observatory API [iNaturalist, 2024; WHO, 2024]) if you would like your app to analyse new data as it is gathered. As part of this process, take note of any common biases or data gaps that are known to exist in these products.

b) <u>*Data Processing*</u>: To reduce unnecessary computation and to streamline your app design, you may wish to include only variables of interest within your data-frame. Depending on the focus of your app, you may also choose to filter your data to focus on key areas, timeframes, species etc (Ambrogi, 2023). Make a thorough note of any changes made during the data processing phase, and make sure to include this in any final reporting.

c) <u>*Selection of LLM parameters*</u>: Research and make decisions on the pre-trained LLM you would like to use for this app. Options include, but are not limited to, Llama, BERT, or the OpenAI GPT models. Make decisions about basic LLM parameters such as scaling temperature (1 = higher probability of more random answers, 0 = more deterministic with low probability of random answers) and verbosity (the length of the generated outputs). Without adding any data or prompts at this stage, use the parameters above to test-run your app.

Phase 2: Prompt engineering and agent evaluation

a) *Create an AI agent to interact with your dataframe:* If using LangChain, create an agent to interact with the SQL or CSV toolkits and attach them to your cleaned data-frame. Test the chatbot to ensure the agent is working correctly and answering user-queries based on the context you provided.

b) *Prompt engineering*: Design meaningful prompts to attach to your agent. This should include a mission statement, metadata descriptions, Q&A examples, and any other meaningful instructions you wish to have attached to every user query.

c) *Iterative testing*: We strongly recommend building a prompt, running your model through a predetermined set of questions, evaluating the correctness and tone of the output, and iteratively evaluating and adjusting your prompts accordingly until a desired threshold for correctness is achieved for your bank of questions. Developers may also consider evaluating the reproducibility of the answers to your test questions.

Phase 3: Application orchestration and deployment

a) *Optional - Multi-agent frameworks*: If you would like to incorporate more source texts into your LLM RAG app, you could build an orchestrated graph app in a system such as LangGraph. We recommend adding a router tool to enable the model to choose between whichever agent deals with the most appropriate text source (e.g. a CSV or SQL agent for matrix data, or a PDF reader for saved literature). Additional tools can be added to evaluate the usefulness of these text sources to the original user query.

b) *Deployment and long-term tracing*: Once you are satisfied with the performance of your app, you can deploy it on a user interface such as Gradio (2024), or Streamlit (2024). Upon deployment, communicate on the user interface (and directly to any relevant stakeholders) that the app provides estimates, and not certainties, to avoid public misunderstanding about the output of the tool. Once the app has been deployed, continue to regularly run audits of its efficacy over time.

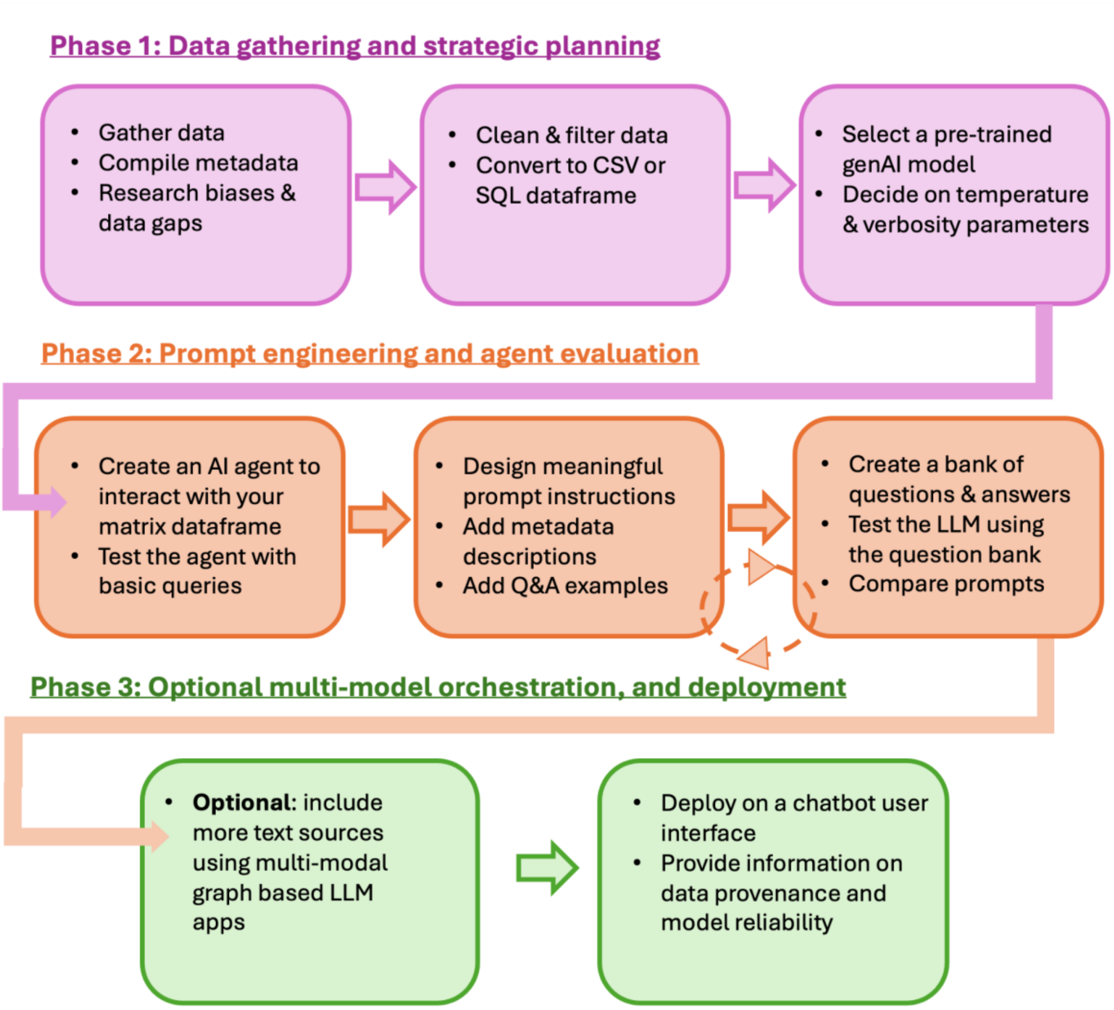

Figure 1: An example workflow for the development, evaluation, and deployment of a retrieval-augmented generation LLM application. Phase 1 (pink) involves gathering your source data and strategically selecting model parameters. Phase 2 (orange) involves designing and iteratively testing prompts and model agents. Phase 3 (green) involves orchestrating multiple LLM agents into a multi-modal graph, and finally deploying the chatbot online.

2. eBird Case Study

We followed our proposed workflow (**Fig.1**) to demonstrate a case-study example of the use of LangChain to build a query-answering framework based on citizen science data. We chose to use eBird data (Sullivan et al., 2014), a global compilation of citizen science bird observations collected by birders, conservationists, and scientists and moderated by ornithology experts. This data can be downloaded at different spatial and temporal resolutions, or imported via API pulls, and contains metadata for each outing, including bird species observed, abundance, sex, breeding or predatory behaviours, exotic status, and space for additional notes by observers. The data contains a mixture of numerical and textual input and therefore provides a useful opportunity to test OpenAI and LangChains' capacity to interpret both qualitative and quantitative data and produce ecologically meaningful LLM output. For this case study, we have focused on an example eBird dataset that was recorded between 1st May - 1st June 2023 in the contiguous counties of Norfolk and Cambridgeshire in the United Kingdom. We analysed the dataset in R (version 4.4.1) to generate a bank of 100 quantitative data interpretation questions (**Appendix Lists 1 and 2**) about bird observations across the study area. We batch-tested these models through the LangChain LLM framework to evaluate whether our models could perform sophisticated reasoning on different types of ecological questions (see **Appendix Table 1**).

We then followed an iterative process of evaluating the LLM outputs against the verified answers and subsequently improving the chain prompt. Our models were iteratively improved over time in line with AI model availability, toolkit production, and enhanced prompts designed to fill in pertinent knowledge gaps as indicated by the previous round of evaluation. The research questions associated with this case study are as follows:

1) Can a chatbot app using LangChain and a pre-trained OpenAI LLM allow us to interact with citizen science matrix data in a scientifically meaningful way?
2) How well does the model perform using different types of ecological query topics?

Through this case study, we aimed to determine whether an LLM can generate accurate and meaningful responses relating to bird abundance and community structure, bird behaviours, and likelihoods of occurrence across different habitat types. In doing so we also investigated whether LLMs are more adept at one aspect of ecological reasoning over another.

Results: eBird Chatbot Case Study

Prompt engineering testing revealed notable improvements to the model when the prompt was iteratively updated (**Fig. 2**). Here, we present the results of each of the seven model variations (see links to the model structure in **Appendix Table 1**). *Model 1* has no prompt, and 46% of the answers fell into the category of ‘Unsure’, whereby the model output stated that this information could not be inferred from the data frame provided. *Model 2* contained a description of basic metadata, including the variable descriptions provided by eBird, and we found a slight decrease in the number of ‘Unsure’ answers (34%) alongside respective increases in answers categorised ‘Correct’ (37%) or ‘Wrong’ (28%). *Model 3* included the same prompt as *Model 2* but with further explanations of variables that were incorrectly interpreted before, nominally variables relating to time and locality, which greatly reduced the number of ‘Unsure’ (17%) answers and increased the number of ‘Correct’ (46%) answers. *Model 4* included the same prompt as *Model 3* but with examples of how questions related to time could be answered (i.e. which column tables to query) and the instruction to try to answer each question to completion and to report ‘I don’t know’ when there was high uncertainty. This model had 62% ‘Correct’, 2% ‘Unsure’ and 36% ‘Wrong’ answers. *Model 5* included the same prompt as *Model 4* but was applied to an eBird dataset of a neighbouring county (Cambridgeshire). The proportion of ‘Correct’ (56%), ‘Unsure’ (7%) and ‘Wrong’ (36%) answers are similar between the initial county dataset (Norfolk) and the Cambridgeshire dataset.

At this stage in our analysis, OpenAI deployed the ChatGPT ‘GPT-4o Mini’ model (OpenAI, 2024) and made it available for developers for use in their own LLM applications. We tested the thorough metadata prompt on ‘GPT-4o Mini’ instead of ‘GPT-3.5’ to form our ‘*Model 6*’. This model performed better than the previous models, with 64.2% ‘Correct’ answers, 6% ‘Unsure’ answers, and 29% ‘Wrong’ answers. Finally, we attached the Wolfram AI tool to the ‘4o Mini’ model with thorough metadata to examine if this tool would enhance the quantitative capacity of the model. For this ‘*Model 7*’, 77% of the answers were ‘Correct’, 2% of the answers were ‘Unsure’, and 21% of the answers were ‘Wrong’.

Across models, ‘*Model 7’* performed best on questions related to bird abundance (e.g. counts of bird observations, n = 25), and questions related to community (e.g. identification of co-occurrence of bird species, n = 39), followed by questions related to the metadata (e.g. questions querying the meaning of the variables, n = 19), and finally questions related to bird behaviour (e.g. questions relating to observer’s fieldwork notes within the data-frame, or question about bird breeding and hunting behaviours, n = 12). For the final version of the model (*Model 7*), 82% of ‘community’ questions were scored as ‘Correct’, 80% of ‘abundance’ questions were scored as ‘Correct’, 68% of ‘metadata’ questions were scored as ‘Correct’, and 67% of ‘behaviour’ questions were scored as ‘Correct’.

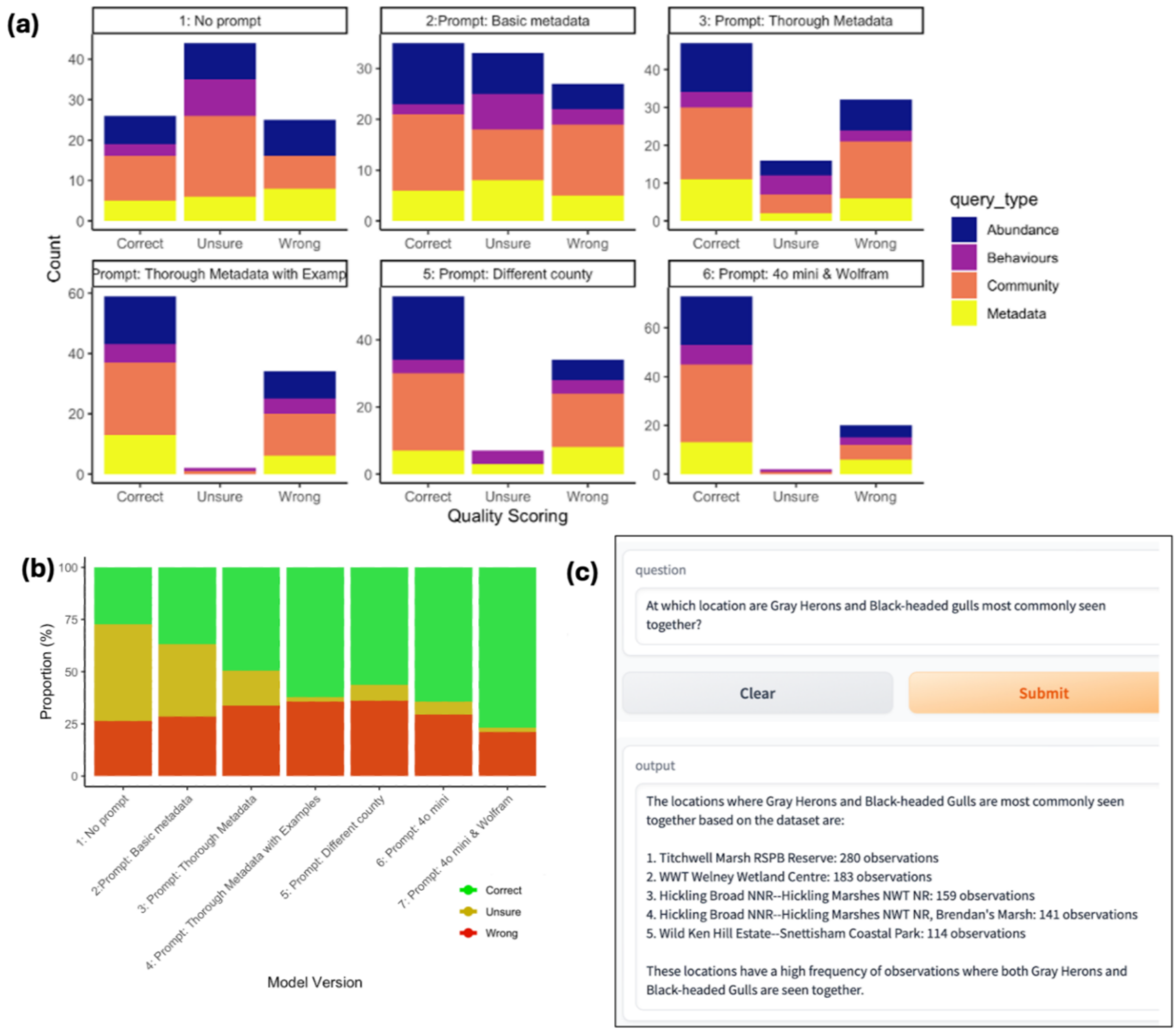

Figure 2: Evaluation of different prompts attached to a LangChain SQL agent. Panel (a) Shows the counts of correct, unsure, and incorrect answers generated by the different model versions, coloured by the different ecological query categories. Panel (b) shows the changing proportion of correct:unsure:incorrect over time as the model was iteratively improved. Panel (c) shows an correct user query and model-generated answer from the final version of the model (version 7), generated in a Gradio user-interface.

<u>**Discussion**</u>

We have devised a proposed workflow for building intelligent, quantitative RAG LLMs adept at interacting with matrix datasets (**Fig.1**). We showcased the design and evaluation procedure for an example model that interacts with citizen science data from eBird (Sullivan et al., 2014), highlighting the importance of iterative prompt design, the use of quantitative agents, and the adaptation to emerging pre-trained LLMs (**Fig.2**). Ultimately, we found that adding detailed metadata descriptions, few-shot examples, and mission statements to the prompts greatly improved model performance and that updates to pre-trained models (e.g. GPT-3.5 to GPT-4o Mini) greatly enhances the ability of the model to interpret user queries, filter, summarise, perform mathematical functions on the data, and produce meaningful answers. The case study model indicates that researchers can customise LLM workflows to create user-friendly tools that interact meaningfully with scientific data. Research to date has focused on the rapidly improving logic and calculus capabilities of pre-trained LLMs (Collins et al., 2024), and the opportunity to design AI agents that can convert plain language queries into mathematical statements and action them in code (Wu et al., 2024). To date, no published LLMs have been trained to carry out more complicated quantitative analyses. However, we predict that these will become widespread as LLMs continue to develop at a rapid rate. For ecologists and environmental scientists working with big data, we recommend keeping abreast of these developments and considering the potential research opportunities that are likely to emerge as a result.

By combining LLMs with existing and robust statistical frameworks, and by using bespoke AI agents and toolkits, we predict that it will soon be possible to create custom RAG systems that can inform real-time conservation, climate adaptation, and public health mitigation actions. Using LLMs instead of traditional statistical tools is innovative and intrinsically scalable. LLMs are adept at understanding and recognising patterns across a diverse array of data types and have already been successfully used to extract useful scientific data in multiple disciplines such as genomics and ecology (Jin et al., 2024; Le Guillarme & Thuiller, 2022; Scheepens et al., 2024). As demonstrated in this paper, LLMs' ability to interpret and analyse structured matrix data using tools like LangChain (particularly the multi-modal LangGraph) offers new possibilities for environmental and ecological research (Topsakal & Akinci, 2023). Data-driven tools could incorporate multi-modal orchestrations (e.g. using LangGraph, see example in **Fig.3**) to draw upon multiple data types, including academic literature, near-real-time matrix data using API pulls, and web-scraping operations. Such tools, if designed carefully and with adequate evaluation (**Fig.1**), could empower policy makers to transform scientific data into actionable interventions at pace.

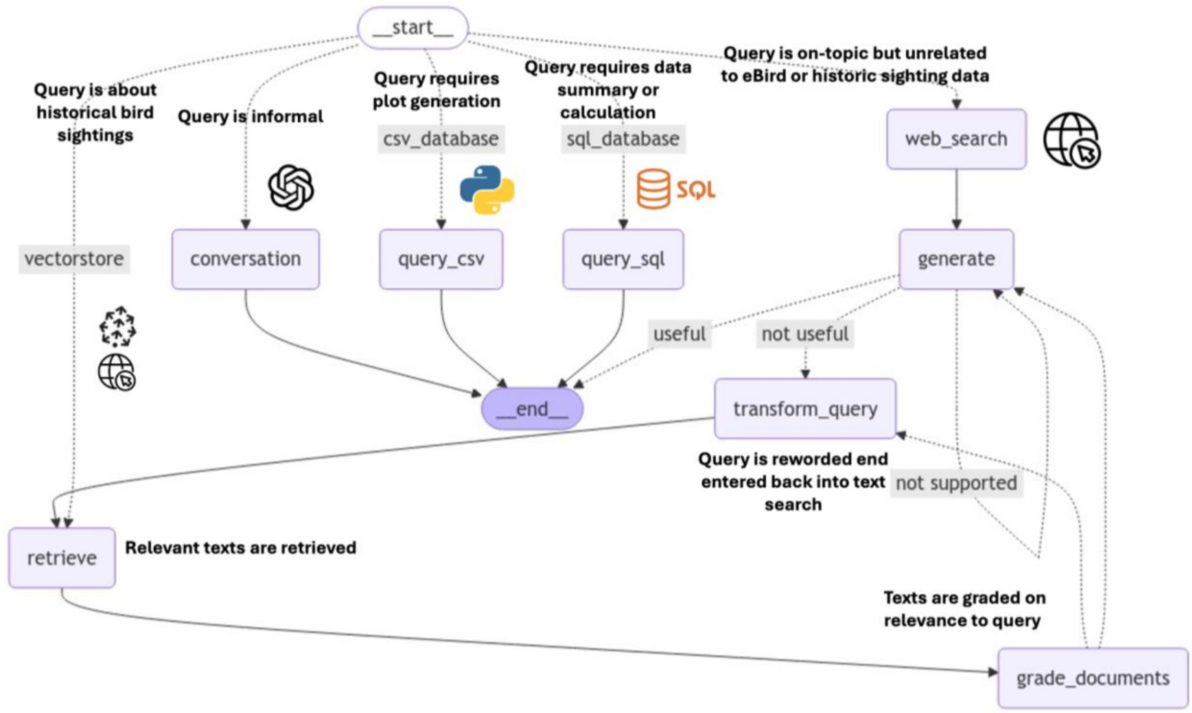

Figure 3: An example multi-modal RAG LLM workflow which incorporates user queries, pre-trained NLP models, custom tools and dataframe agents and multiple data sources, with a variety of visual, textual and numerical outputs. Model outline built in LangGraph.

One clear benefit of integrating LLMs into the analysis of ecological data is the increased timeliness of response time between initial data collection and data-informed action (Marvin et al., 2016). Camera-trapping and audio monitoring are increasingly becoming enhanced by AI neural network technology, bridging the gap between *in situ* data monitoring and species identification and geolocation (Wall et al., 2008; Ware et al., 2012). Likewise, by pairing quantitative LLMs with near-real-time environmental data (e.g. OpenWeather), and citizen science data, AI technology could save critical data cleaning and analysis time for quantitative ecologists by allowing them to outsource more menial data activities and focus instead on scientific inquiry, stakeholder collaboration, and outreach (Lamba et al., 2019; McClure et al., 2020). Furthermore, integrating LLMs and citizen science data may boost engagement between the public (particularly citizen science contributors) and science, especially if the gap between data publication and analysis is facilitated by AI frameworks (Pecl et al., 2019; Theobald et al., 2015). Accessible AI tools can also promote communication across research and policy sectors, making it easier to transform raw ecological data into action. The rapid uptake of neural network technology in the sphere of ecological research (McClure et al., 2020; Torney et al., 2019; Willi et al., 2019) indicates that researchers are willing to explore the analysis capabilities of other AI tools as and when they develop (Christin et al., 2019). It is therefore important to build and uphold robust and sustainable development and evaluation frameworks for these tools.

We recommend that quantitative researchers building RAG LLMs consider the concept of "garbage in, garbage out" when choosing the data to include within their model, to the same extent one would when building a traditional statistical framework (Kilkenny & Robinson, 2018). As with any quantitative analysis, the quality of the output is contingent on the quality of the data provided to the LLM. Ecological monitoring data can be prone to issues of selective bias towards charismatic species, misidentification, and inclusion of data entry errors. For example, GBIF data has high degrees of spatial bias, which in turn can skew the results of species distribution models (Beck et al., 2014). Furthermore, citizen science databases which are compiled by non-expert observers, can be messy, biassed by site selection, weather conditions, and selective observation of particular species and behaviours (Dobson et al., 2020; Thornhill et al., 2016; Tulloch et al., 2013). Researchers can adjust their statistical model designs to reflect such biases, for example through standardising observation counts between sites and building multilevel hierarchical models (Bird et al., 2014). However, these data transformation methods may be less reliably actioned using LLM agents alone. We recommend that any vital data processing and preparation is conducted before quantitative analysis is performed by a RAG LLM (**Fig.1**; Phase 1).

Pre-trained AI models update at a high frequency, though at a cost to reproducibility for developers building upon these base models (Ma et al., 2024). We experienced such a shift ourselves during the testing of our eBird case study model, whereby 'GPT-4o-Mini' was introduced towards the end of our investigation - helpfully highlighting both the iterative improvements of new LLM releases, and also the rapid pace of development (**Fig.2**). We predict that the high deprecation rate of LLM releases will remain high as their capabilities are tested, and that any prospective developers keep abreast of new updates. In designing our roadmap for building and evaluating LLM apps (**Fig.1**), we aimed to frame our suggestions broadly enough that they may be applied across new and unforeseen software developments. Another common issue faced by developers using pre-trained LLMs is the high level of stochasticity and non-determinism of results when the model temperatures are higher, and that the "black box" nature of pre-trained LLMs can make transparency, reproducibility, and quality-testing difficult (Ceccaroni et al., 2019; McClure et al., 2020; Ollion et al., 2024; Ouyang et al., 2024). These issues highlight the need to *a)* design thorough prompts which ask your model to report its logic when generating an answer, and *b)* ensure that the deployed version of your LLM apps clearly state that the model is AI and has the propensity to make mistakes (**Fig.1**; Phases 2 & 3).

Conclusion

There is strong potential to enhance the accessibility, speed, and effectiveness of ecological and environmental data analysis through the development of quantitative RAG LLMs. By integrating advanced, pre-trained AI LLMs with existing ecological and environmental data, ecologists can build customisable 'virtual statisticians' that streamline data analysis, making trend detection and actionable insights more readily available and fast-tracking the route from data collection through to communication to policy-makers. Through our demonstration of the eBird chatbot, we show how researchers can integrate AI tools to empower them to ask nuanced questions about biodiversity patterns and trends. Ecologists may wish to take advantage of the emerging research capabilities of AI, but we urge them to do so with an awareness of the risks inherent across LLM models. We have provided a roadmap for developing multimodal LLM apps responsibly and transparently, while leveraging ongoing model updates. As AI technologies continue to advance, the opportunities to bridge the gap between data collection and data-driven interventions will proliferate. LLM innovations may be the key to transforming raw data into rapid insights that drive ecological and environmental solutions. It is therefore the responsibility of ecologists now to develop, promote, and pursue sustainable AI research frameworks in order to guide the future of responsible and impactful science.

Data availability:

Scripts and data used to conduct the eBird chatbot case study are available for review and download at: BioDivHealth/eBird_testing: Testing the model output of eBird LLMs

Reference List

Abdelmageed, N., Löffler, F., & König-Ries, B. (2023). BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain. In SWAT4HCLS (pp. 62-71).

Ambrogi, M. (2023, September 18). *10 Ways to Improve the Performance of Retrieval Augmented Generation Systems*. Medium. https://towardsdatascience.com/10-ways-to-improve-the-performance-of-retrieval-augmented-generation-systems-5fa2cee7cd5c

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology Letters*, *15*(4), 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Ceccaroni, L., Bibby, J., Roger, E., Flemons, P., Michael, K., Fagan, L., & Oliver, J. L. (2019). Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. *Citizen Science: Theory and Practice*, *4*(1), Article 1.

Ceccaroni, L., Oliver, J. L., Roger, E., Bibby, J., Flemons, P., Michael, K., & Joly, A. (2023). *Advancing the productivity of science with citizen science and artificial intelligence*. https://doi.org/10.1787/69563b12-en

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

Chatterjee, S., Bhattacharya, M., Lee, S.-S., & Chakraborty, C. (2023). Can artificial intelligence-strengthened ChatGPT or other large language models transform nucleic acid research? *Molecular Therapy Nucleic Acids*, *33*, 205–207. https://doi.org/10.1016/j.omtn.2023.06.019

Chen, H., & Ding, Y. (2025, January). Implementing traffic agent based on LangGraph. In Fourth International Conference on Intelligent Traffic Systems and Smart City (ITSSC 2024) (Vol. 13422, pp. 582-587). SPIE.

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), 1632–1644. https://doi.org/10.1111/2041-210X.13256

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, *121*(24), e2318124121. https://doi.org/10.1073/pnas.2318124121

Dobson, A. D. M., Milner-Gulland, E. J., Aebischer, N. J., Beale, C. M., Brozovic, R., Coals, P., Critchlow, R., Dancer, A., Greve, M., Hinsley, A., Ibbett, H., Johnston, A., Kuiper, T., Le Comber, S., Mahood, S. P., Moore, J. F., Nilsen, E. B., Pocock, M. J. O., Quinn, A., … Keane, A. (2020). Making Messy Data Work for

Conservation. *One Earth*, *2*(5), 455–465. https://doi.org/10.1016/j.oneear.2020.04.012

Doney, S. C., Ruckelshaus, M., Duffy, J. E., Barry, J. P., Chan, F., English, C. A., Galindo, H. M., Grebmeier, J. M., Hollowed, A. B., Knowlton, N., Polovina, J., Rabalais, N. N., Sydeman, W. J., & Talley, L. D. (2012). Climate Change Impacts on Marine Ecosystems. *Annual Review of Marine Science*, *4*(Volume 4, 2012), 11–37. https://doi.org/10.1146/annurev-marine-041911-111611

Enríquez-de-Salamanca, Á. (2025). Botanical databases in EIA: opportunities and challenges. Impact Assessment and Project Appraisal, 1-11.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Google. (2024). *Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processin*. http://research.google/blog/open-sourcing-bert-state-of-the-art-pre-training-for-natural-language-processing/

Gougherty, A. V., & Clipp, H. L. (2024). Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *Npj Biodiversity*, *3*(1), 1–5. https://doi.org/10.1038/s44185-024-00043-9

Gradio. (2024). *Gradio*. https://www.gradio.app/

Haag, C., Steinemann, N., Chiavi, D., Kamm, C. P., Sieber, C., Manjaly, Z.-M., Horváth, G., Ajdacic-Gross, V., Puhan, M. A., & Wyl, V. von. (2023). Blending citizen science with natural language processing and machine learning: Understanding the experience of living with multiple sclerosis. *PLOS Digital Health*, *2*(8), e0000305. https://doi.org/10.1371/journal.pdig.0000305

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., ... & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. Nature, 637(8045), 319-326.

iNaturalist. (2024). *A Community for Naturalists · iNaturalist*. https://www.inaturalist.org/

Jeong, C. (2024). *A Study on the Implementation Method of an Agent-Based Advanced RAG System Using Graph* (arXiv:2407.19994). arXiv. https://doi.org/10.48550/arXiv.2407.19994

Jin, Q., Yang, Y., Chen, Q., & Lu, Z. (2024). GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, *40*(2), btae075. https://doi.org/10.1093/bioinformatics/btae075

Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: "Garbage in – garbage out". *Health Information Management Journal*, *47*(3), 103–105. https://doi.org/10.1177/1833358318774357

Lamba, A., Cassey, P., Segaran, R. R., & Koh, L. P. (2019). Deep learning for environmental conservation. *Current Biology*, *29*(19), R977–R982. https://doi.org/10.1016/j.cub.2019.08.016

LangChain. (2024). *LangChain*. https://www.langchain.com/

LangGraph. (2024). *LangGraph*. https://www.langchain.com/langgraph

Le Guillarme, N., & Thuiller, W. (2022). TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, *13*(3), 625–641. https://doi.org/10.1111/2041-210X.13778

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-

Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C. and Dai, D., 2024. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.

Ma, Z., Mei, Y., Gajos, K. Z., & Arawjo, I. (2024). Schrödinger's Update: User Perceptions of Uncertainties in Proprietary Large Language Model Updates. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3613905.3651100

Marvin, D. C., Koh, L. P., Lynam, A. J., Wich, S., Davies, A. B., Krishnamurthy, R., Stokes, E., Starkey, R., & Asner, G. P. (2016). Integrating technologies for scalable ecology and conservation. *Global Ecology and Conservation*, *7*, 262–275. https://doi.org/10.1016/j.gecco.2016.07.002

McClure, E. C., Sievers, M., Brown, C. J., Buelow, C. A., Ditria, E. M., Hayes, M. A., Pearson, R. M., Tulloch, V. J. D., Unsworth, R. K. F., & Connolly, R. M. (2020). Artificial Intelligence Meets Citizen Science to Supercharge Ecological Monitoring. *Patterns*, *1*(7). https://doi.org/10.1016/j.patter.2020.100109

Ollion, É., Shen, R., Macanovic, A., & Chatelain, A. (2024). The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*, *6*(1), 4–5. https://doi.org/10.1038/s42256-023-00783-6

OpenAI. (2024). *Introducing ChatGPT | OpenAI*. https://openai.com/index/chatgpt/

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2024). *An Empirical Study of the Non-determinism of ChatGPT in Code Generation* (arXiv:2308.02828). arXiv. https://doi.org/10.48550/arXiv.2308.02828

Pecl, G. T., Stuart-Smith, J., Walsh, P., Bray, D. J., Kusetic, M., Burgess, M., Frusher, S. D., Gledhill, D. C., George, O., Jackson, G., Keane, J., Martin, V. Y., Nursey-Bray, M., Pender, A., Robinson, L. M., Rowling, K., Sheaves, M., & Moltschaniwskyj, N. (2019). Redmap Australia: Challenges and Successes With a Large-Scale Citizen Science-Based Approach to Ecological Monitoring and Community Engagement on Climate Change. *Frontiers in Marine Science*, *6*. https://doi.org/10.3389/fmars.2019.00349

Pollock, L. J., Kitzes, J., Beery, S., Gaynor, K. M., Jarzyna, M. A., Mac Aodha, O., ... & Berger-Wolf, T. (2025). Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. Nature Reviews Biodiversity, 1-17.

Scheepens, D., Millard, J., Farrell, M., & Newbold, T. (2024). Large language models help facilitate the automated synthesis of information on potential pest controllers. *Methods in Ecology and Evolution*, *15*(7), 1261–1273. https://doi.org/10.1111/2041-210X.14341

Streamlit. (2024). *Streamlit • A faster way to build and share data apps*. https://streamlit.io/

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M. A., & Parrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, *181*, 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

Thornhill, I., Loiselle, S., Lind, K., & Ophof, D. (2016). The Citizen Science Opportunity for Researchers and Agencies. *BioScience*, *66*(9), 720–721. https://doi.org/10.1093/biosci/biw089

Topsakal, O., & Akinci, T. C. (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *International Conference on Applied Engineering and Natural Sciences*, *1*, 1050–1056. https://doi.org/10.59287/icaens.1127

Torney, C. J., Lloyd-Jones, D. J., Chevallier, M., Moyer, D. C., Maliti, H. T., Mwita, M., Kohi, E. M., & Hopcraft, G. C. (2019). A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, *10*(6), 779–787. https://doi.org/10.1111/2041-210X.13165

Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J., & Martin, T. G. (2013). Realising the full potential of citizen science monitoring programs. *Biological Conservation*, *165*, 128–138. https://doi.org/10.1016/j.biocon.2013.05.025

Wall, D. H., Bradford, M. A., ST. JOHN, M. G., Trofymow, J. A., Behan-Pelletier, V., Bignell, D. E., Dangerfield, J. M., Parton, W. J., Rusek, J., & Voigt, W. (2008).

Global decomposition experiment shows soil animal impacts on decomposition are climate-dependent. *Global Change Biology*, *14*(11), 2661–2677.

Ware, C., Bergstrom, D. M., Müller, E., & Alsos, I. G. (2012). Humans introduce viable seeds to the Arctic on footwear. *Biological Invasions*, *14*(3), 567–577. https://doi.org/10.1007/s10530-011-0098-4

WHO. (2024). *Global Health Observatory*. https://www.who.int/data/gho

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

Willis, K. J., & Bhagwat, S. A. (2009). Biodiversity and Climate Change. *Science*, *326*(5954), 806–807. https://doi.org/10.1126/science.1178838

Wolfram Research, Inc., Mathematica, Version 14.1, Champaign, IL (2024).

Wu, Y., Jia, F., Zhang, S., Li, H., Zhu, E., Wang, Y., Lee, Y. T., Peng, R., Wu, Q., & Wang, C. (2024, March 15). *MathChat: Converse to Tackle Challenging Math Problems with LLM Agents*. ICLR 2024 Workshop on Large Language Model (LLM) Agents. https://openreview.net/forum?id=S7vIB7OGQe

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). *Large Language Models Are Human-Level Prompt Engineers* (arXiv:2211.01910). arXiv. https://doi.org/10.48550/arXiv.2211.01910

Appendix

Appendix List 1: Batch questions for eBird chatbot evaluation, using data from the county of Norfolk (United Kingdom)

Question
What is the total count of birds observed in Norfolk?
Provide the average duration of surveys in Norfolk.
How many surveys were conducted at Cley & Salthouse Marshes?
How many unique bird species have been observed at Titchwell.
What is the total count of Bearded Reedling observed in Norfolk?
How many surveys recorded sightings of the Western Marsh Harrier in Norfolk?
What is the average number of bird species observed per survey at Blakeney?
Provide the total sum of Black Headed Gull observed at NWT Holme Dunes.
How many observations were recorded at Happisburgh?
How many observations of exotic birds were there?
Which is the most common bird observed?
Which is the least common bird observed?
What are the 3 most common birds at Hardwick Flood Lagoon?
Rank the top 10 most common bird species observed at Cromer Golf Course.
What is the total count of Carrion Crows observed in Norfolk?
What is the total number of Carrion Crows by observation count ?
What are the 3 most common birds in Norfolk?
What are the 3 most common birds by observation count?
What is the average number of bird species observed per survey at Stiffkey Fen?
Provide the total count of Manx Shearwaters observed at Sidestrand.
How many surveys were conducted on 5th May?
How many observations were there on 5th May?
How many unique species were seen on 5th May?
How many exotic bird species are observed at Stiffkey Fen.
What is the total count of birds observed at Wensum Park in the first week of May?
Provide the average duration of observer effort per survey at Cringleford Marsh.

How many surveys conducted at Cringleford Marsh included Coal Tit?
List all the bird species observed at Cromer Golf Course before 5pm on May 5th.
What is the least common bird species observed at Stiffkey?
Where was the most northerly sighting of the Smew?
What is the average number of bird species observed per survey at Stiffkey Fen during May?
What was the median duration effort?
Which location has the highest number of individual observers?
Which bird is most often seen stationary?
What is the scientific name for the Purple Heron?
Where are Common Buzzards more abundant, at Stiffkey Fen or at Titchwell Marsh?
Which species are never spotted at Cromer Golf Course but are spotted elsewhere?
Which two species are most likely to be observed together in the same survey?
Which three species are most likely to be observed together in the same survey at Cromer Golf Course?
Which two species are most likely to be observed together in the same survey at Stiffkey?
Which bird species can only be observed before 8am at Sidestrand?
Which bird species can only be observed after 2pm at Sidestrand?
What is the third most abundant bird species at Titchwell Marsh?
Which bird species are only observed once at Snettisham RSPB Reserve?
Where are Barnacle Geese more likely to be observed, at Stiffkey Fen or at Titchwell Marsh?
Which bird species are most likely to be observed in a stationary position at Titchwell Marsh?
What is the scientific name for the Little Egret?
Where are Black-bellied Plovers more abundant, at Stiffkey Fen or at Holme Dunes?
At which locality am I most likely to see a Ruddy Shelduck (by sightings)?
At which locality am I most likely to see a Ruddy Shelduck (by abundance)?
Which two species are most likely to be observed together in the same survey at Titchwell Marsh?

Which bird species can only be seen after 8pm at Holme Dunes?
Are more birds observed in the morning or in the afternoon?
Are more bird species observed in the morning or in the afternoon?
What is the second rarest bird species at Whitlingham Country Park?
Where are Gray Herons more likely to be observed, at Holme Dunes or at Whitlingham Country Park?
Which observer has the highest count of Western Marsh Harrier observations?
Which observer has the lowest count of Western Marsh Harrier observations?
Are Gray Herons more likely to be seen in the same surveys as Black-headed Gulls or Eurasian Wrens?
At which location are Gray Herons and Black-headed gulls most commonly seen together?
Which two species are most commonly observed together at Cromer Golf Course?
Which two species are least commonly observed together at Cromer Golf Course?
Which location has the least temporally consistent data?
What is the scientific name of the most abundant bird species at Holme Dunes?
Which bird species is usually observed earliest every day at Titchwell Marsh?
Where are *Charadrius hiaticulas* more likely to be observed, at Stiffkey Fen or at Holme Dunes?
What is the average number of Garganeys spotted per observation?

What species have not been reported at Happisburgh but could potentially be found there?

Which species are rarely observed but have been spotted in high numbers when seen?

Which species are more commonly observed but have been spotted in low numbers when seen?
Were any unusual behaviours noted among the birds in the dataset?
Was any breeding activity observed at Titchwell Marsh?
Which bird species was seen with most breeding activity at Titchwell Marsh?
Were any hunting behaviours observed in Norfolk?
How do birds interact with their environment?
Which bird species are not seen in the northwest of Norfolk but are seen elsewhere?

What sensitive species were observed in Norfolk?
Which is the southernmost species of plover?
Which species are always solitary?
At which location are the largest groups of the same bird species observed?
Which location has the highest diversity of birds?
Which location has the lowest diversity of birds?
Which birds have been seen on rainy days?
Were any predatory behaviors observed at Holme Dunes?
How do birds in the community at Cromer Golf Course interact with their environment?
Which 5 bird species are most commonly seen in fens?
What reasons are given for unapproved observations?
What species are most commonly reported by group surveys?
Which bird species are most likely to have species notes?
Find the observer with the highest average number of species per checklist
Were any unusual behaviours noted among the birds at Holme Dunes?
What time of day am I most likely to spot a Smew?
Which birds are seen on rainy days at Dersingham Bog NNR?
Which birds are overrepresented?
Are weekdays or weekends better for observing different types of birds?

Appendix List 2: Batch questions for eBird chatbot evaluation, using data from the county of Cambridgeshire (United Kingdom)

Question
What is the total count of birds observed in Cambridgeshire?
Provide the average duration of surveys in Cambridgeshire.
How many surveys were conducted at Grafham Water?
How many unique bird species have been observed at Smithy Fen.
What is the total count of Bearded Reedling observed in Cambridgeshire?
How many surveys recorded sightings of the Western Marsh Harrier in Cambridgeshire?
What is the average number of bird species observed per survey at Wicken Fen NNR?
Provide the total sum of Black Headed Gull observed at Roswell Pits.
How many observations were recorded at Cambridge Botanic Garden?
How many observations of exotic birds were there?
Which is the most common bird observed?
Which is the least common bird observed?
What are the 3 most common birds at Coe Fen?
Rank the top 10 most common bird species observed at Grantchester Meadows.
What is the total count of Carrion Crows observed in Cambridgeshire?
What is the total number of Carrion Crows by observation count ?
What are the 3 most common birds in Cambridgeshire?
What are the 3 most common birds by observation count?
What is the average number of bird species observed per survey at Grafham Water?
Provide the total count of Arctic Tern observed at Fen Drayton Lakes RSPB Reserve.
How many surveys were conducted on 5th May?
How many observations were there on 5th May?
How many unique species were seen on 5th May?
How many exotic bird species were observed at Grafham Water.
What is the total count of birds observed at Dernford Reservoir in the first week of May?
Provide the average duration of observer effort per survey at Paradise LNR.
How many surveys conducted at Paradise LNR included Mallard?

List all the bird species observed at Grantchester Meadows before 5pm on May 10th.
What is the least common bird species observed at Grafham Water?
Where was the most northerly sighting of the Barn Owl?
What is the average number of bird species observed per survey at Grafham Water during May?
What was the median duration effort?
Which location has the highest number of individual observers?
Which bird is most often seen stationary?
What is the scientific name for the Barn Owl?
Where are Common Buzzards more abundant, at Grafham Water or at Smithy Fen?
Which species are never spotted at Grantchester Meadows but are spotted elsewhere?
Which two species are most likely to be observed together in the same survey?
Which three species are most likely to be observed together in the same survey at Grantchester Meadows?
Which two species are most likely to be observed together in the same survey at Wicken Fen?
Which bird species can only be observed before 8am at Fen Drayton Lakes RSPB Reserve?
Which bird species can only be observed after 2pm at Fen Drayton Lakes RSPB Reserve?
What is the third most abundant bird species at Smithy Fen?
Which bird species are only observed once at Emmanuel College?
Where are Mallards more likely to be observed, at Grafham Water or at Smithy Fen?
Which bird species are most likely to be observed in a stationary position at Smithy Fen?
What is the scientific name for the Little Egret?
Where are Common Chiffchaffs more abundant, at Grafham Water or at Roswell Pits?
At which locality am I most likely to see a Ruddy Shelduck (by sightings)?
At which locality am I most likely to see a Ruddy Shelduck (by abundance)?
Which two species are most likely to be observed together in the same survey at Emmanuel College?
Can a Dunnock be seen after 4pm at Emmanuel College?
Are more birds observed in the morning or in the afternoon?
Are more bird species observed in the morning or in the afternoon?

What is the second rarest bird species at Coe Fen?
Where are Gray Herons more likely to be observed, at Wicken Fen or at Coe Fen?
Which observer has the highest count of Western Marsh Harrier observations?
Which observer has the lowest count of Western Marsh Harrier observations?
Are Gray Herons more likely to be seen in the same surveys as Black-headed Gulls or Eurasian Wrens?
At which location are Gray Herons and Black-headed gulls most commonly seen together?
Which two species are most commonly observed together at Grantchester Meadows?
Which two species are least commonly observed together at Grantchester Meadows?
Which location has the least temporally consistent data?
What is the scientific name of the most abundant bird species at Wicken Fen?
Which bird species is usually observed earliest every day at Smithy Fen?
Where are *Anas platyrhynchos* more likely to be observed, at Grafham Water or at Roswell Pits?
What is the average number of Garganeys spotted per observation?
What species have not been reported at Cambridge Botanic Garden but could potentially be found there?
Which species are rarely observed but have been spotted in high numbers when seen?
Which species are more commonly observed but have been spotted in low numbers when seen?
Were any unusual behaviours noted among the birds in the dataset?
Was any breeding activity observed at Fen Drayton?
Which bird species was seen with the most breeding activity at Fen Drayton?
Were any hunting behaviours observed in Cambridgeshire?
How do birds interact with their environment?
Which bird species are not seen in the northwest of Cambridgeshire but are seen elsewhere?
What sensitive species were observed in Cambridgeshire?
Which is the southernmost species of plover?
Which species are always solitary?
At which location was the largest group of the same bird species observed?

Which location has the highest diversity of birds?
Which location has the lowest diversity of birds?
Which birds have been seen on cloudy days?
Were any predatory behaviors observed at Roswell Pits?
How do birds in the community at Grantchester Meadows interact with their environment?
Which 5 bird species are most commonly seen in fens?
What reasons are given for unapproved observations?
What species are most commonly reported by group surveys?
Which bird species are most likely to have species notes?
Find the observer with the highest average number of species per checklist
Were any unusual behaviours noted among the birds at Roswell Pits?
What time of day am I most likely to spot a Mallard?
Which birds are seen on cloudy days at Paradise LNR?
Which birds are overrepresented?
Are weekdays or weekends better for observing different types of birds?

Appendix Table 1: A summary table of the prompts used in the testing of the eBird Langchain chatbot. All full prompts and additional code can be found: https://github.com/BioDivHealth/eBird_testing/tree/main/scripts/python

Model #	OpenAI GPT Model used	Description	Github Link
1	3.5-turbo	No prompt	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/Python/1_basic_ebird_dashboard.py
2	3.5-turbo	Column metadata from eBird	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/python/2_metadata_prompt_ebird_dashboard.py
3	3.5-turbo	Column metadata from eBird + explanations of ecological concepts (e.g. abundance)	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/python/3_thorough_metadata_prompt_ebird_dashboard.py
4	3.5-turbo	Column metadata from eBird + explanations of ecological concepts + example responses (e.g. how to interpret time variables)	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/python/4_thorough_metadata_examples_prompt_ebird_dashboard.py
5	3.5-turbo	Column metadata from eBird explanations of ecological concepts + example responses **[Cambridgeshire data and questions from Appendix List 2]**	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/python/5_camb_throrough_metadata_examples_prompt_ebird_dashboard.py
6	4o-mini	Column metadata from eBird explanations of ecological concepts + example responses	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/python/6_4o_habitatprompt_ebird.py
7	4o-mini	Column metadata from eBird explanations of ecological concepts + example responses + Wolfram Alpha LangChain Tool	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/python/7_wolfram_4o_habitatprompt_ebird.py
8	4o-mini	Column metadata from eBird explanations of ecological concepts + example responses **[and additional data sources, using a work-in-progress LangGraph orchestration]**	https://github.com/BioDivHealth/eBird_testing/blob/main/scripts/python/8_langgraph_mutli_agent_rag_sql_graphing.py