Received: 17 June 2025 | Accepted: 4 October 2025

DOI: 10.1111/2041-210X.70184

Methods in Ecology and Evolution

PERSPECTIVE

Fast-tracking ecological interpretation using bespoke quantitative large language models

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, UK

[2]UK Centre for Ecology and Hydrology, Penicuik, UK

Correspondence
Elise C. Gallois
Email: elise.gallois94@gmail.com

Funding information
Natural History Museum Science Investment Fund; Sir Henry Dale Research Fellowship, Grant/Award Number: 20179/Z/20/Z and 220179/A/20/Z

Handling Editor: Edward Codling

Abstract

1. The Anthropocene presents significant challenges for global biodiversity, public health and 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 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 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. To build robust 'virtual quantitative assistants' 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 discuss potential use-case examples that apply the LangChain framework to analyse citizen science data. Using our LLM roadmap, we highlight the importance of iterative and strategic prompt engineering and agent selection, in addition to iteratively evaluating model output. As LLM software continues to evolve, its integration into ecological and environmental research can empower ecologists with purpose-built tools that bridge the gap between data collection and actionable solutions.

KEYWORDS
artificial intelligence, citizen science, large language models, multi-agent models, natural language processing

1 | INTRODUCTION

The Anthropocene offers 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 rapidly, tools to investigate and communicate emerging phenomena within these datasets remain time-consuming and specialised. Artificial intelligence (AI) data tools could allow a facilitation 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; Liu et al., 2024; OpenAI, 2024; Topsakal & Akinci, 2023). While this expansive evolution may pose challenges for long-term reproducibility, it also presents emerging opportunities for scientific efficiency. For instance, LLMs have been used to auto-generate patient discharge forms based on 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 metadata about pathogen hosts (Gougherty & Clipp, 2024).

There is untapped potential to use NLPs 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 (Hollmann et al., 2025). 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 publicly available quantitative ecological data, LLMs could enable us to effectively communicate with our datasets, fast-track data interpretation and facilitate 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 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 policymakers.

In this paper, we present a roadmap for developing custom multi-modal LLMs to serve as virtual data assistants—or 'virtual quantitative assistants'—designed to support ecologists in summarising, visualising and exploring trends within complex ecological datasets. These tools represent a timely opportunity for ecological researchers to interact with data in more intuitive and accessible ways. 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. 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.

2 | METHODS

2.1 | 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 Figure 1 and is then used to showcase the design, implementation and evaluation of a working citizen science chatbot. 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.

2.1.1 | 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

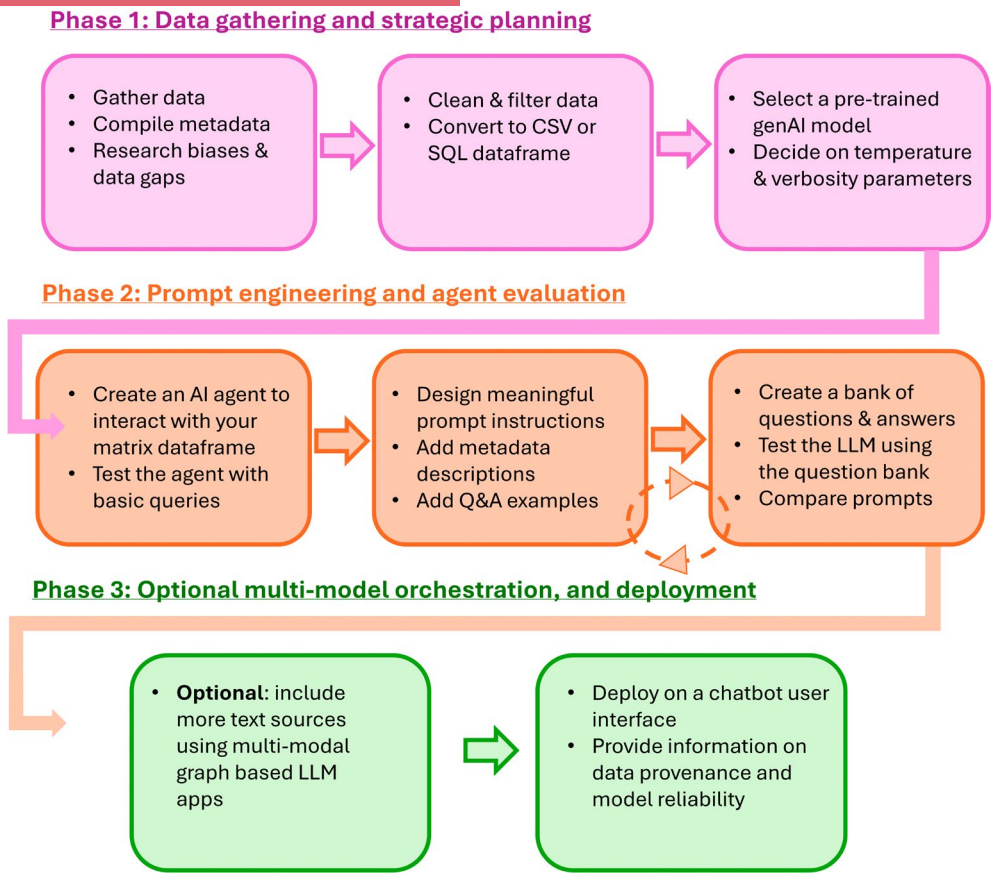

FIGURE 1 An example workflow for the development, evaluation and deployment of a retrieval-augmented generation large language model (LLM) application. Phase 1 involves gathering your source data and selecting model parameters. Phase 2 involves designing and testing prompts and agents. Phase 3 involves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

that allow developers to connect these models with bespoke data sources and to create interactive systems. Retrieval augmented generation (RAG) models combine pretrained 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 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 operation workflows (LangChain, 2024) and designing intelligent urban traffic control tools (Chen & Ding, 2025).

2.1.2 | 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 keywords, or to pay particular attention to data features (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 subquestions 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 towards 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; Figure 1) and are integral components to bespoke and advanced RAG models.

2.1.3 | 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 (e.g. producing plots) and are executed when selected by an LLM 'agent' which acts as a decision-making component that reads the user input and predesigned prompt and routes the query to the appropriate tool (Jeong, 2024; LangChain, 2024; Topsakal & Akinci, 2023). A suite of toolkits and agents exists 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 WolframAlpha tool can connect LLM chains to the WolframAlpha computational search engine to facilitate the computation of more complex mathematical tasks (Wolfram Research, Inc, 2024). We recommend using agents and tools for summarising, visualising and performing mathematical operations on ecological and environmental matrix data within RAG LLM models. R is a commonly used language in the field of ecology, and Python workflows can also be embedded into R sessions using the 'reticulate' interface (R Core Team, 2021; Ushey et al., 2025). 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.

2.1.4 | 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 a 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) or an open-source alternative such as LLMLayer (2025) or the DuckDuckGo (2021) LangChain tool. 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.

2.1.5 | 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 full roadmap, split into three phases, for the development of such an app (Figure 1).

Phase 1: Data gathering and strategic planning

a. *Data gathering*: Select the quantitative dataframe you would like to provide as the key data source for your app. If available, collate all relevant metadata explaining data provenance (e.g. eBird citation), variable names (e.g. observation counts) and units. You may choose a static dataframe to upload manually to your coding environment or 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. *Data processing*: To reduce unnecessary computation and to streamline your app design, you may wish to include only variables of interest within your dataframe. 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). Thoroughly document changes made during the data processing phase and included this in any final reporting.
c. *Selection of LLM parameters*: Research and make decisions on the pretrained 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 dataframe. 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 (i.e. few-shot prompting, whereby examples of inputs and expected outputs are provided to demonstrate

expected responses from the model, Brown et al., 2020), and any other meaningful instructions you wish to have attached to every user query.

c. *Iterative testing*: Users should build a prompt, run their model through a predetermined set of questions, evaluate the correctness and tone of the output and iteratively evaluate and adjust their prompts accordingly until a desired threshold for correctness is achieved for the bank of questions. For example, a bank of 100 questions could be generated and answers precalculated using non-AI-assisted analyses. These questions should be asked in every iteration of the model, and evaluators could score each answer for accuracy, helpfulness and tonal appropriateness. These scores can then be visualised and statistically evaluated to track the development of model performance as improvements are iteratively made. 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. Router tools can 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 UI (and directly to any stakeholders) that the app provides estimates and not certainties to avoid public misunderstanding about the tool's outputs. Once the app has been deployed, continue to regularly run audits of its efficacy over time.

2.2 | eBird case study

We followed our proposed workflow to demonstrate a case study example of the use of LangChain to build a query-answering framework based on citizen science data. We used 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. These data can be downloaded at different spatial and temporal resolutions and contain metadata for each outing, including bird species observed, abundance, sex, breeding or predatory behaviours, exotic status and space for additional observer notes. The data contain both numerical and textual input and therefore provide a useful opportunity to test OpenAI and LangChain's capacity to interpret both qualitative and quantitative data and produce ecologically meaningful LLM output. In Appendix S1, we outline the methodology for the design and evaluation of this case study application, including a bank of evaluation questions (Appendix S1: Lists 1 and 2) and full model comparison results for different development stages of the model (Appendix S1: Table A1). To test model performance against different types of questions an ecological researcher may wish to pose, we categorised the 100 questions into four query categories: bird abundance, community ecology, metadata interpretation and bird behaviour. The research questions associated with this case study are as follows:

1. Can a chatbot app using LangChain and a pretrained 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?

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 interpretation over another. All data and code used in this analysis are available to download (Gallois, 2025).

3 | RESULTS: eBird CHATBOT CASE STUDY

Prompt engineering testing revealed notable improvements to the model when the prompt was iteratively updated (Figure 2). We present the results of each of the seven model variations in detail in Appendix S1, whereby we present the changing proportions of 'Correct', 'Unsure' and 'Wrong' answers (see links to the model structure in Appendix S1: Table A1 and exact model prompts in the available codebase). We started with a model with no additional prompt (Model 1), 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 dataframe provided. The models provided greater accuracy when prompts were iteratively refined (Model 2: basic metadata added to prompt, Model 3: detailed metadata and variable descriptions added to prompt and Model 4: thorough prompt with example question-and-answer pairs and further clarification for variables related to time), and when calculus agents were added to the model workflow and when newer versions of OpenAI models were integrated (Models 6 and 7). The final model version, with thorough metadata in the prompt, OpenAI 4o Mini and the Wolfram tool included, had 77% 'Correct' answers, 2% 'Unsure' answers and 36% 'Wrong' answers. We categorised our question bank into four types of ecological query: bird abundance, community dynamics, metadata interpretation and bird behaviour. The models did not perform equally well across types of ecological query, with higher accuracy scores reported for user questions relating to bird abundance and community composition compared to user questions relating to metadata reporting and bird behaviour. 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',

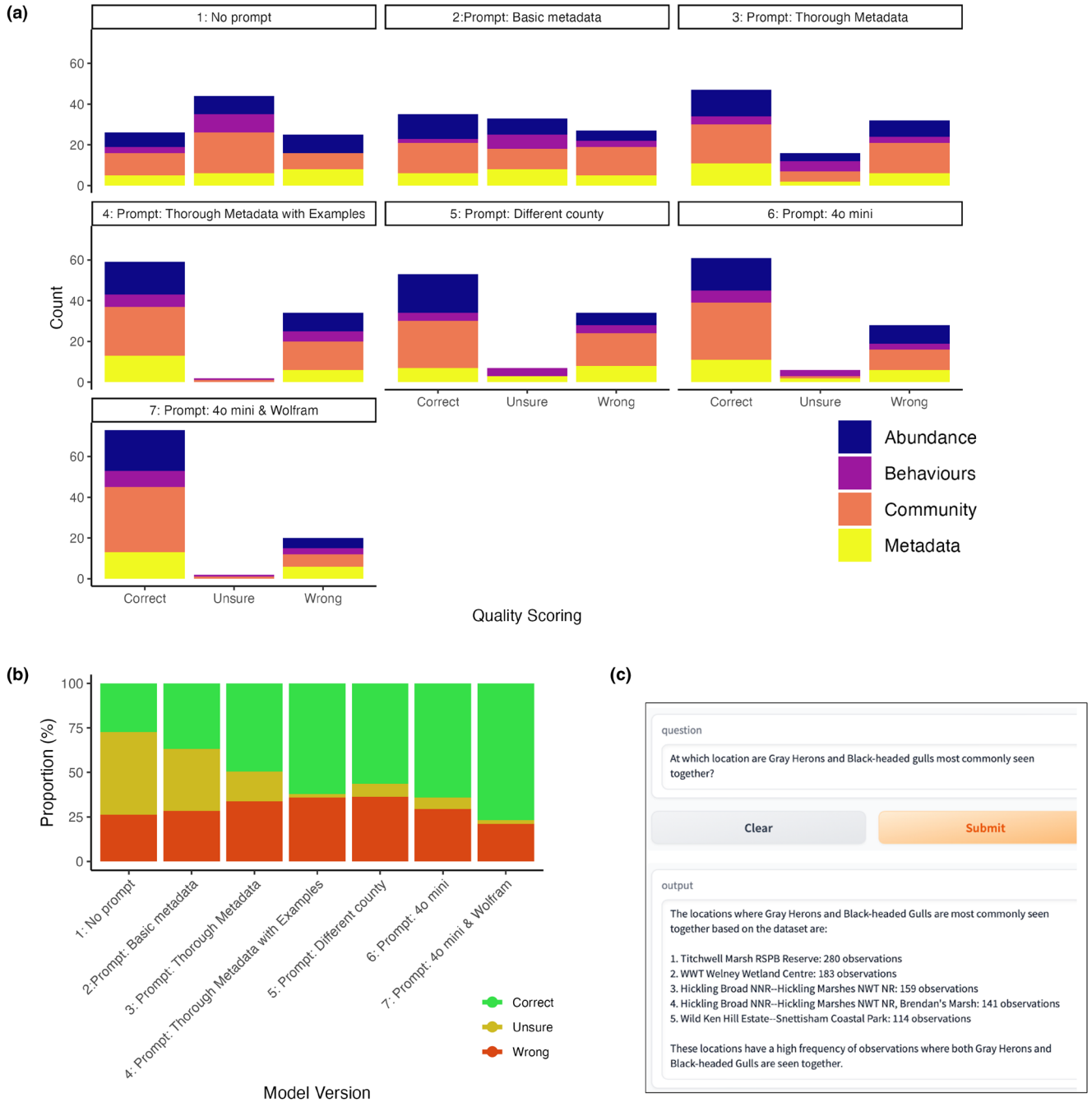

FIGURE 2 Evaluation of different prompts attached to a LangChain SQL agent. (a) The counts of correct, unsure and incorrect answers generated by the different models, coloured by the different ecological query categories. (b)The changing proportion of correct:unsure:incorrect over time as the model was iteratively improved. (c) A correct user query and model-generated answer from the final version of the model, visualised in a Gradio UI.

68% of 'metadata' questions were scored as 'Correct' and 67% of 'behaviour' questions were scored as 'Correct'.

4 | DISCUSSION

We have devised a proposed workflow for building intelligent, quantitative RAG LLMs adept at interacting with matrix datasets (Figure 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 pretrained LLMs (Appendix S1; Figure 2). Adding detailed metadata descriptions, few shot examples and mission statements to the prompts improved the ability of the model to interpret user queries, filter, summarise, perform mathematical functions on the data and produce meaningful answers. Users can use the chatbot to generate outputs which accessibly translate dense and complex datasets

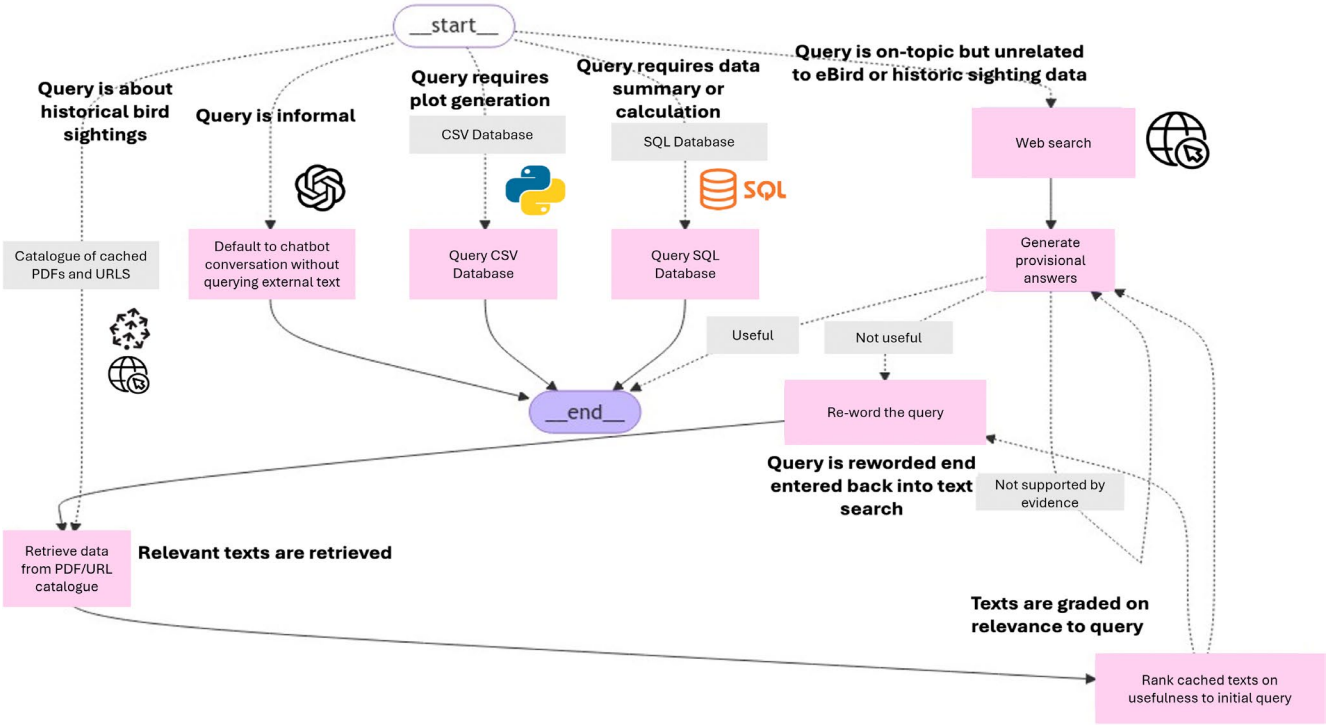

FIGURE 3 An example multi-modal retrieval augmented generation large language model workflow which incorporates user queries, pretrained natural language processing models, custom tools and dataframe agents and multiple data sources, with a variety of visual, textual and numerical outputs.

into accessible, plain language statements. However, our case study also revealed that quantitative LLM tools do not necessarily produce equally useful interpretations for all types of ecological queries. For example, our models are more likely to be adept at questions requiring data summarisation or questions about the geography and timing of bird observations and less adept at questions requiring a deeper understanding of animal behaviour such as hunting, mating and migration. Our findings also revealed the rapid improvements in model performance because of the rollout of a new LLM model version (i.e. OpenAI 4o mini). Our case study provides an example of how a researcher can iteratively improve their LLM 'helper' tool, while also highlighting existing blind spots in LLMs.

Research to date has focused on the rapidly improving logic and calculus capabilities of pretrained 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. Bottlenecks may be encountered if researchers need to devote additional time to learning AI methods alongside their existing workloads, particularly with a lack of existing guidance about model design, quality assurance and output evaluation. 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. In this regard, we encourage researchers to incorporate frameworks such as ours into their AI-supplemented workflows to ensure best practice and minimise time-consuming mistakes.

By combining LLMs with existing and robust statistical frameworks, and by using bespoke AI agents and toolkits, it is becoming possible to create scalable custom RAG systems that can inform real-time conservation, climate adaptation (e.g. Juhasz et al., 2024) and public health actions (e.g. Ng et al., 2025). LLMs can identify and understand patterns across a diverse array of data types and have already been successfully used to extract useful scientific data in disciplines such as genomics and ecology (Jin et al., 2024; Le Guillarme & Thuiller, 2022; Scheepens et al., 2024). 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 Figure 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, could empower policymakers to transform scientific data into actionable interventions at pace.

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 and citizen science data, AI technology could help reduce repetitive data wrangling tasks and accelerate early-stage analyses, enabling quantitative ecologists to devote more time to model design, interpretation and broader scientific inquiry, while also facilitating collaboration and generating or editing content for further outreach (Lamba et al., 2019; McClure et al., 2020). Furthermore, integrating LLMs and citizen science data may boost engagement between the public (particularly data 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 promote communication across research and policy sectors by helping transform raw ecological data into actionable insights, but their outputs should undergo transparent quality assurance and control checks (i.e. for factual accuracy, bias reduction and tonal appropriateness) by domain experts prior to (and following) the public deployment of any LLM tool. 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 input. 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 have 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, biased 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 non-inferential analysis is performed by LLMs (Figure 1; Phase 1), and critically that human researchers lead the design and implementation of any inferential statistics.

Pretrained 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 (Figure 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 (Figure 1), we aimed to frame our suggestions broadly enough that they may be applied across new and unforeseen software developments. Many of the most popular pretrained LLM tools (such as OpenAI GPTs—Anthropic, 2024; Google, 2025) are not open source, with source code privately secured by developers. Users of these pretrained LLMs typically must purchase low-cost tokens to run these tools as part of their own code and are at risk of models being permanently pulled by developers. Another common issue faced by developers using pretrained 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 pretrained 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 (Figure 1, Phases 2 and 3).

Although our framework indicates that LLMs can interact successfully with quantitative data, we do not advocate that these tools 'replace' quantitative ecologists, who are critically needed for designing and implementing rigorous statistical modelling. Unlike other AI tools, such as neural networks or image classifiers, LLMs are designed to understand and reproduce language patterns and have not been explicitly trained for inferential analyses at this stage. Our pilot case study suggests that LLMs can be adapted to query and summarise quantitative data. LLMs can also visualise trends in datasets and communicate numerical output in plain English. Quantitative ecologists are essential for identifying meaningful knowledge gaps and designing statistical models, interpreting results and ensuring good practice and quality control throughout scientific inquiry. We see LLMs as complementary virtual assistants, which can help reduce data interpretation bottlenecks and workloads if integrated into a researcher's workflow. We also acknowledge that unnecessary or excessive use of resource-intensive AI carries its own environmental footprint (Dhar, 2020; Pollock et al., 2025), and ecologists should weigh these risks carefully against potential benefits. The role of ecologists is therefore not replaced by LLMs. Instead, ecologists may now include critical testing of AI inputs and contextualising LLM outputs as part of their work.

5 | 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, pretrained 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 policymakers. 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 to guide the future of responsible and impactful science.

AUTHOR CONTRIBUTIONS

Elise C. Gallois and David W. Redding conceived of this project together. Arianna Salili-James, Sanson T. S. Poon and Artur Trebski provided coding and conceptual assistance to Elise C. Gallois in the coding and design of the conceptual roadmap. Elise C. Gallois wrote the manuscript with feedback from all co-authors.

ACKNOWLEDGEMENTS

We would like to thank the NHM AI Lab Programme (https://doi.org/10.3897/biss.8.138147) for supporting this research.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70184.

DATA AVAILABILITY STATEMENT

Data available at https://doi.org/10.5281/zenodo.15319206 (Gallois, 2025).

ORCID

Elise C. Gallois https://orcid.org/0000-0002-9402-1931
Arianna Salili-James https://orcid.org/0000-0003-1125-2054
Sanson T. S. Poon https://orcid.org/0000-0001-5297-7452
Artur Trebski https://orcid.org/0009-0006-3259-5215
David W. Redding https://orcid.org/0000-0001-8615-1798

REFERENCES

Abdelmageed, N., Löffler, F., & König-Ries, B. (2023). *BiodivBERT: A pre-trained language model for the biodiversity domain*. SWAT4HCLS.

Ambrogi, M. (2023). *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

Anthropic. (2024). *Claude Haiku [large language model]*. https://www.anthropic.com

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

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), 29.

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). Implementing traffic agent based on LangGraph. In H. Chen & W. Shangguan (Eds.), *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 of the United States of America*, *121*(24), e2318124121. https://doi.org/10.1073/pnas.2318124121

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, *2*(8), 423–425.

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*, 11–37. https://doi.org/10.1146/annurev-marine-041911-111611

DuckDuckGo. (2021). *LangChain*. Langchain.com. https://python.langchain.com/docs/integrations/tools/ddg/

Enríquez-de-Salamanca, Á. (2025). Botanical databases in EIA: Opportunities and challenges. *Impact Assessment and Project Appraisal*, *43*, 1–312.

Gallois, E. (2025). Leveraging large language models for ecological interpretation using an eBird chatbot case study [data set]. *Zenodo*. https://doi.org/10.5281/zenodo.15319206

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 (arXiv:2312.10997)*. 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/

Google. (2025). *Google Gemini [large language model]*. https://gemini.google.com/

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., & von Wyl, V. (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 (arXiv:2407.19994)*. 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

Juhasz, M., Dutia, K., Franks, H., Delahunty, C., Mills, P. F., & Pim, H. (2024). Responsible retrieval augmented generation for climate decision making from documents. *arXiv preprint arXiv:2410.23902*. https://doi.org/10.48550/arXiv.2410.23902

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., & Dai, D. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*. https://doi.org/10.48550/arXiv.2412.19437

LLMLayer. (2025). *LLMLayer*. https://llmlayer.ai/

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), 100109. https://doi.org/10.1016/j.patter.2020.100109

Ng, K. K. Y., Matsuba, I., & Zhang, P. C. (2025). RAG in health care: A novel framework for improving communication and decision-making by addressing LLM limitations. *Nejm AI*, *2*(1), AIra2400380.

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 (arXiv:2308.02828)*. 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*, 349. https://doi.org/10.3389/fmars.2019.00349

Pollock, L. J., Kitzes, J., Beery, S., Gaynor, K. M., Jarzyna, M. A., Mac Aodha, O., Meyer, B., Rolnick, D., Taylor, G. W., Tuia, D., & Berger-Wolf, T. (2025). Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nature Reviews Biodiversity*, *1*, 166–182.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Ushey, K., Allaire, J., & Tang, Y. (2025). *reticulate: Interface to 'Python'*. R package Version 1.43.0. https://rstudio.github.io/reticulate/

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. (2024). *Mathematica, Version 14.1*.

Wu, Y., Jia, F., Zhang, S., Li, H., Zhu, E., Wang, Y., Lee, Y. T., Peng, R., Wu, Q., & Wang, C. (2024). *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 (arXiv:2211.01910)*. https://doi.org/10.48550/arXiv.2211.01910

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Methodology and results write-up for the design and evaluation of the eBird chatbot.

How to cite this article: Gallois, E. C., Salili-James, A., Poon, S. T. S., Trebski, A., & Redding, D. W. (2025). Fast-tracking ecological interpretation using bespoke quantitative large language models. *Methods in Ecology and Evolution*, *16*, 2730–2740. https://doi.org/10.1111/2041-210X.70184