Large language models possess some ecological knowledge, but how much?

Filip Dorm[1,2] | Joseph Millard[3] | Drew Purves[4] | Michael Harfoot[5] | Oisin Mac Aodha[1]

[1]School of Informatics, University of Edinburgh, UK

[2]School of Architecture, Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[3]Department of Zoology, University of Cambridge, UK

[4]Google DeepMind, UK

[5]Vizzuality, Cambridge, UK

Correspondence
Oisin Mac Aodha
Email: oisin.macaodha@ed.ac.uk

Funding information
FD and OMA were supported by a Royal Society Research Grant. JM was funded by the Leverhulme Trust and the Isaac Newton Trust on an Early Career Fellowship.

Abstract

1. Large Language Models (LLMs) have garnered attention for their potential at question answering across multiple different domains. However, despite the promise demonstrated, there has only been limited exploration of their effectiveness in the context of ecological knowledge.
2. We investigate the ecological knowledge and potential reasoning abilities of two LLMs, `Gemini 1.5 Pro` and `GPT-4o`, across a suite of ecologically focused tasks. Our tasks quantitatively assess a models' ability to predict species presence at locations, generate range maps, list critically endangered species, classify threats, and estimate species traits. We compare model performance against expert-derived data to quantify their accuracy and reliability using a new benchmark dataset we introduce.
3. We show that while the LLMs tested outperform naive baselines, they still exhibit significant limitations, particularly in generating spatially accurate range maps and classifying threats.
4. Our findings underscore both the potential and challenges of using LLMs in ecological applications, highlighting the need for further refinement, including domain-specific fine-tuning, to better approximate ecological reasoning. Our new benchmark dataset will enable researchers to make progress on this task by providing a repeatable way to evaluate future models.

KEYWORDS
Ecological Question Answering, Evaluation, LLMs

1 | INTRODUCTION

The biodiversity crisis is one of the defining challenges of our time, with species loss accelerating at an unprecedented rate (Barnosky et al., 2011). As ecologists strive to fill the ever-growing gaps in biodiversity knowledge (Hortal et al., 2015), novel approaches to data collection and analysis are becoming increasingly critical. In this context, Large

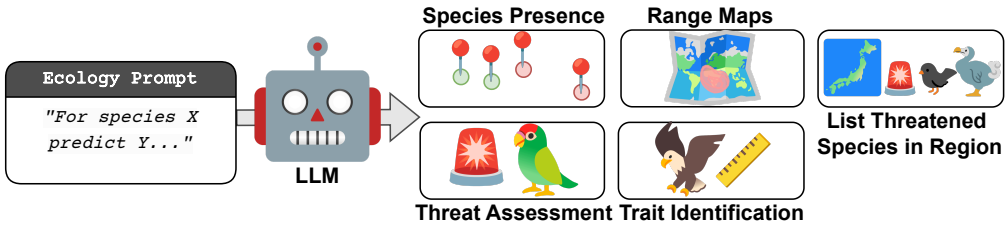

FIGURE 1 **Evaluation Overview**. We quantitatively evaluate existing off-the-shelf large language models (LLMs) on a set of five different ecological question answering tasks. The tasks require the LLMs to answer questions related to identifying what species are present at a specified location, mapping their ranges, estimating species' traits, assessing specific threat levels, and listing threatened species.

Language Models (LLMs), trained on vast amounts of textual data, have the potential to be a transformative tool for ecology. However, while current LLMs have shown utility in domains ranging from coding and mathematics to literature (Poldrack et al., 2023; Frieder et al., 2023; Mirzadeh et al., 2024; Chang et al., 2023), their application to ecological research remains largely unexplored. Could these models assist in processing ecological data to identify species at risk, predict species distributions, or inform conservation decision making? This study seeks to answer these questions by evaluating the ecological information recall of state-of-the-art LLMs through a series of carefully designed tasks.

Several studies have already explored how LLMs can process spatial information, which is important for ecological applications such as modeling species distributions (Zhang et al., 2024; Roberts et al., 2024b; Tao and Xu, 2023; Deng et al., 2024; Manvi et al., 2023, 2024). Particularly relevant is GPT4GEO (Roberts et al., 2024a), whose evaluation tasks include estimating the spatial distribution of species and migratory routes of birds. In spatial assessments by LLMs, studies such as (Manvi et al., 2023), suggest that LLMs are better at making predictions for locations described by place names, likely because these align with the textual format of the training data the models have previously encountered. LLMs have also been utilized for species distribution modeling, enabling the use of text information describing species' ranges and habitats when modeling their distributions (Sastry et al., 2023; Hamilton et al., 2024).

While LLMs have been recognized for their ability to assist with tasks such as extracting information from scientific literature and coding (Doi et al., 2024; Gougherty and Clipp, 2024; Cooper et al., 2024; Farrell et al., 2024; Atuhurra et al., 2024; Scheepens et al., 2024), ensuring the accuracy of their predictions remains crucial. Addressing challenges related to erroneous outputs, (Mora-Cross and Calderon-Ramirez, 2024) explored methods for estimating uncertainty in LLM-generated results, thereby mitigating risks in ecological applications. LLMs also show promise in debunking sensational wildlife news stories (Santangeli et al., 2024) and assessing invasive species distributions based on news reports (Castro et al., 2024).

Recently, LLMs have acquired multi-modal capabilities, allowing for joint reasoning over images and text (Miao et al., 2024). For example, (Fabian et al., 2023) assessed how such multi-modal models can be used to automate the annotation of wildlife imagery. Similar to how (Maruf et al., 2024) assess Vision Language Models (VLMs) on a defined benchmark for trait classification, we aim to complement such a benchmark by focusing on the capabilities of language models.

Previous evaluation of language-only models for predicting trait-related quantities primarily focused on extracting information from provided datasets and text (Domazetoski, 2024; Marcos et al., 2024). In contrast, we demonstrate that LLMs can generate this information independently, without relying on additional data. For example, while previ-

ous research like (Le Guillarme and Thuiller, 2022) specializes in taxonomic entity recognition, our findings reveal that LLMs can effectively perform similar tasks directly, eliminating the need for supplementary datasets.

We present a suite of species-centric evaluation tasks to assess LLMs' factual knowledge and reasoning in ecology (see Figure 1). Our tasks are quantifiable against datasets, ensuring scalability and reproducibility without relying on subjective evaluation. We employ off-the-shelf models, emphasizing the importance of identifying their limitations, such as hallucinations, to ensure safe use in ecology. We evaluate two state-of-the-art LLMs, `GPT-4o` from OpenAI (ChatGPT, 2024) and `Gemini 1.5 Pro` from Google (Gemini, 2024), through a series of experiments testing their performance in the ecological domain. https://github.com/filipgdorm/eco-llm contains the code for reproducing our experiments.

2 | MATERIAL AND METHODS

2.1 | Models

Our experiments consist of querying two commercial LLMs, `Gemini 1.5 Pro` and `GPT-4o`, through their publicly accessible APIs. We selected these two models as they are representative of the best commercial LLMs currently available. Both models were used in the period between August and October 2024. For both models, the temperature, a hyperparameter that controls the randomness of a model, was set to zero to ensure deterministic outputs, thereby enhancing the reproducibility of the results. However, the results may still vary when there are changes or updates to the underlying models themselves. Importantly, neither `Gemini 1.5 Pro` nor `GPT-4o` retains conversation history when utilizing the API, which is advantageous for our purposes, as we do not want the model to learn from previous queries.

Gemini 1.5 Pro. We used `gemini-1.5-pro-001`, as at the time of writing, it is the most powerful version of the Gemini model available. This model was queried through its API (Gemini - API, 2024).

GPT-4o. We used `GPT-4o`, the exact model version `gpt-4o-2024-05-13`, as it is, at the time of writing, the most powerful version they have. This model was queried through its API (GPT-4 - API, 2024).

2.2 | Datasets

Spatial Distribution Data. The species presence and range map tasks are based on expert-derived range maps of 2,418 species from the International Union for Conservation of Nature (IUCN) Red List (IUCN, 2024a). This is the same evaluation dataset used in existing work for evaluating species distribution models (Cole et al., 2023; Dorm et al., 2024). For each species in the dataset, each cell location on the earth of size 250km^2 is marked as containing the species or not. Compared to presence-only crowdsourced location data, this presence-absence data offers a more complete representation of the species' distributions, as it includes explicit absence information in addition to recorded presences, making it more suitable for our evaluation.

Threat Data. For the threat classification task, we query the IUCN's API (IUCN Red List - API, 2024) to extract the Red List status of species, countries they reside in, and threats they face.

Trait Data. For tasks related to predicting numerical traits for species, we utilized two primary data sources. For bird traits, we used AVONET (Tobias et al., 2022), and for mammals, we used COMBINE (Soria et al., 2021). Both datasets encompass a wide range of species with average measurements of morphological traits.

TABLE 1 **Task 1: Species Presence**. Species Presence. We present mean accuracy (with scores binarized at 0.5) and mean average precision (MAP) for 100 species, along with the count of unparseable model responses.

Method	Task 1a - Locations			Task 1b - Countries		
	Mean Acc.	MAP	#Invalid R.	Mean Acc.	MAP	#Invalid R.
`Gemini 1.5 Pro`	70.8%	76.9%	14	84.9%	**86.9%**	2
`GPT-4o`	**78.0%**	**77.5%**	0	**87.6%**	85.1%	1
Random Baseline	51.6%	52.2%	-	51.3%	49.3%	-
Always Absent Baseline	50.0%	50.0%	-	68.3%	31.7%	-
Always Present Baseline	50.0%	50.0%	-	31.7%	31.7%	-

2.3 | Evaluation Tasks

We designed a set of evaluation tasks to test the ecological information recall of LLMs. These five tasks allow us to test an LLMs' knowledge of the geographical distribution of species, their ability to assess threats, and understand species' size and morphology.

3 | RESULTS

3.1 | Task 1: Species Presence

Task 1 Prompt

```
Can you tell me if the <common_name> (<scientific_name>) can be found at the
following locations:<[Location 1, ... Location 10]>. Please respond with a score
for each location between 0 and 1, where 1 indicates that you think the species
is very likely to be present there. Provide your answer as a Python list
containing only the likelihood scores.
```

FIGURE 2 **Prompt template for Task 1: Species Presence**.

The first task focuses on identifying species presence or absence across ten specific locations, aiming to evaluate LLMs' understanding of species' geographical distributions (Task 1a). For this task, we used 100 species for which range maps were available. These 100 species were randomly selected to ensure an equal representation of the four taxonomic classes in the dataset: 25 amphibians, 25 reptiles, 25 mammals, and 25 birds. Here, each species was paired with ten coordinate locations, with five indicating presence and five absence. These locations are sampled from IUCN expert range maps. To add complexity, two of the absence locations are "difficult", indicating that they are sampled from areas adjacent to the species' known range. Specifically, locations are represented as the center points of H3 hexagons (H3, 2024) at resolution five, which cover an average area of 253 km². Difficult absences are sampled from resolution five hexagons within larger resolution three parent cells (12,393 km²) of the known range where the species is not expected to occur. This approach ensures that the absence locations are close to, but not within, the species' known range.

The experiment was repeated using country names instead of coordinates (Task 1b). In this case, the number of

presence countries varied from one (for species inhabiting only a single country) to a maximum of five. Absences were drawn from neighboring countries where the species does not occur. Random countries were sampled as absences if there were not enough neighboring countries to have ten countries for the prompt. The same prompt structures were used but the locations were replaced with countries.

Both models frequently provided supplementary details in their responses, such as the reasoning behind their predictions or disclaimers regarding their capabilities. A response was deemed valid if it contained identifiable opening and closing brackets of a Python list, with the content in between being loadable without errors. Evaluation metrics were computed from the valid and parseable responses. As shown in Table 1 and Figure 10, both `Gemini 1.5 Pro` and `GPT-4o` significantly outperform the naïve baselines, which include randomly guessing probabilities, always predicting presence (i.e., 1), or always predicting absence (0). In Figure 10, we visualize the task and predictions on a map for a subset of ten species.

`GPT-4o` outperforms `Gemini 1.5 Pro` in identifying expert-derived presences, as shown by the mean accuracy and MAP metrics in Table 1. This performance advantage is also reflected in the number of invalid outputs, with `GPT-4` producing fewer of these. Additionally, both LLMs perform better when asked to assess whether the species are present in countries rather than at coordinate locations, as seen in Table 1. This is to be expected as the model is more likely to have encountered this information before rather than providing a random coordinate, which requires more sophisticated pattern recognition capabilities.

3.2 | Task 2: Range Maps

Task 2 Prompt

```
My goal is to create a GEOJSON describing the geographical distribution of the
<common_name> (<scientific_name>). I am looking for a correctly formatted,
parsable GEOJSON file. Please provide a complete GEOJSON that I can then render
and nothing else.
```

FIGURE 3 **Prompt template for Task 2: Range Maps.**

This task involves asking the LLMs to generate range maps for the same 100 species as in Task 1. Assessing LLMs' geographic awareness through their range map accuracy provides valuable insights into their ability to handle spatial data.

For this task, we ask the model to return a GEOJSON string, as these objects are parseable and can be visualized easily. The overlap with the expert range maps and the F1 score can be computed from this, as the F1 score measures how well a model balances precision (the proportion of correctly predicted presences among all predicted presences) and recall (the proportion of correctly predicted presences among all actual presences), refer to the Appendix B for more details. The GEOJSON is extracted by locating the curly braces {} in the response, and the enclosed string, including the braces, is then parsed.

The mean F1 score of the LLM-generated range maps for the 100 evaluation species is reported in Table 2, while the predicted range maps for ten of these species are visualized in Figure 4. `Gemini 1.5 Pro` generates more accurate range maps on average, as indicated by its mean F1 score being four percentage points higher. However, it also produces a greater number of non-parseable maps. In most cases, both models are successful in generating range maps that overlap with the true ranges of the species, as indicated by their non-zero F1 scores for the majority of

TABLE 2 **Task 2: Range Maps**. Here we report the mean F1 score of the range maps generated by different LLMs compared with the expert-derived maps. In addition, we also compare to a recent deep learning-based species distribution model SINR (Cole et al., 2023) (specifically using a coordinate only SINR model with the $\mathcal{L}_{\text{AN-full}}$ loss, capped at 1000 training samples per species and without environmental covariates) using either a fixed threshold of 0.1 or a threshold estimated using the LPT-R approach from (Dorm et al., 2024).

Range Map Creation Method	Mean F1 Score	# Invalid R.
`Gemini 1.5 Pro`	24.7%	26
`GPT-4o`	20.9%	15
SINR (threshold 0.1)	54.9%	-
SINR (threshold LPT-R)	60.7%	-

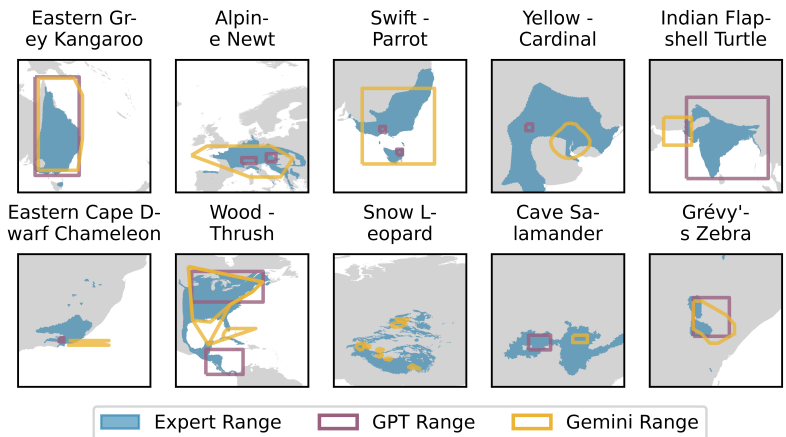

FIGURE 4 **Task 2: Range Maps**. Here we display the range maps generated by `Gemini 1.5 Pro` (yellow) and `GPT-4o` (purple) as well as the expert-derived range maps from IUCN (blue) we use for evaluation.

species. The most common reason for the model's failure to generate a range map was instances where it produced excessively long lists of boundary coordinates, resulting in either a violation of response length rules or a loss of coherence in the output; a notable example of this is the Snow Leopard. Out of the range maps provided in Figure 4, `Gemini 1.5 Pro` generates range maps at a finer resolution, while `GPT-4o` simply draws a square. Of these ten species, `Gemini 1.5 Pro` generates three range maps with an F1 score above 60%, namely: Eastern Grey Kangaroo, Alpine Newt, and Wood Thrush, while `GPT-4o` only reaches this performance for Eastern Grey Kangaroo. However, despite these maps achieving a higher F1 score, they still exhibit low spatial resolution.

We attempted to specify in the prompt that we were looking for approximately 50 coordinates for the border outline to generate higher-resolution range maps that were less square-like. However, the LLMs produced predictions with lower accuracy and more frequent parsing errors, such as missing commas and improperly formatted coordinates in the output. We opted to use GeoJSON because the model was more consistent in generating a predictable and parseable output format, adhering to its structure more strictly than when simply outputting a list of border coordinates.

In addition, we also compare our results to a recent deep learning-based species distribution model, SINR (Cole

et al., 2023). For this comparison, we utilized a coordinate-only version of the SINR model with the $\mathcal{L}_{\text{AN-full}}$ loss function, capped at 1,000 training samples per species and without incorporating any environmental covariates. Predictions were made using either a fixed threshold of 0.1 or a threshold estimated through the LPT-R approach as outlined in (Dorm et al., 2024). This comparison shows that the range maps produced by LLMs are not as accurate as those produced by other deep-learning-based methods.

3.3 | Task 3: List Threatened Species in Region

Task 3 Prompt

```
List the birds that are classified as Critically Endangered on the IUCN Red List
and are found in <country>.

Instructions:
- Provide the species names using their common names only.
- Format your response as a Python list, like this: [Species name 1, Species name
2, ...].
- Ensure that there are no duplicate species names in your list.
```

FIGURE 5 **Prompt template for Task 3: Threatened Species.**

This task involves evaluating the LLMs' ability to comprehensively list species that meet specific criteria for a particular geographic region. While this assessment focuses on identifying 'critically endangered' species, it allows us to gauge the LLMs' effectiveness as a tool for researchers and conservationists in identifying species that may be at risk in a given area.

The task aimed to evaluate the ability of LLMs to retrieve and compile information about species that belong to the most severe threat class of the IUCN (i.e., critically endangered—CR) within specific geographic regions. The experiment was intentionally designed to incorporate a geographic element to test how well the models can synthesize information across multiple dimensions (species, threat status, and location) based on their training data. While feeding a set of species and querying their threat status independently could isolate the threat retrieval component, our goal was to assess the combined capability of the LLM to identify relevant species given a location-based constraint. This approach aligns with real-world conservation use cases, where the geographic scope in assessing species health is often constrained to specific regions, such as evaluating the conservation status of species within protected areas, national parks, or even entire countries—common scenarios in initiatives like the IUCN Red List assessments or regional conservation programs. The CR category includes only species that have been formally assessed, meaning it does not account for other species that may also be at risk of extinction but have not yet undergone assessment. For our purposes, however, this classification was sufficient to evaluate how well the LLM can retrieve and compile information based on its training data. We chose this threat class to ensure that it is theoretically possible for LLMs, given their response length limitations, to list all species in a single answer - Indonesia has the highest number of CR birds with 30 unique species in that class. The prompt used is depicted in Figure 5.

When interpreting the results in Figure 6, it is important to recognize that the task inherently combines geographic and threat status dimensions. This design was intentional to evaluate the LLM's ability to handle this complexity. The data itself is based on political boundaries, meaning the results reflect jurisdictional classifications. For example, both Congo and Lithuania report no critically endangered species according to IUCN records. Similarly, regions like

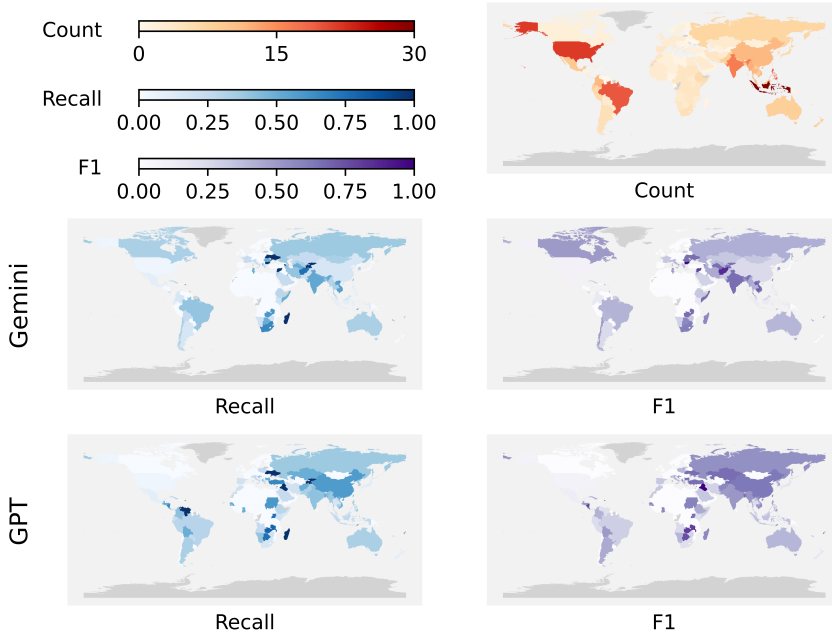

FIGURE 6 **Task 3: Threatened Species**. (Top) The number of critically endangered (CR) bird species across different countries, with darker red shades representing higher numbers. Grey indicates regions where data is unavailable. (Middle) The recall and F1 of `Gemini 1.5 Pro` when asked to list CR birds in different countries, with darker blue and purple shades representing higher recall and F1 respectively. (Bottom) The recall and F1 of `GPT-4o`.

Greenland may show zero recorded species, as they are likely classified under Denmark's jurisdiction in the dataset. Countries with no available data or those not recognized as independent entities in our plotting tool are marked in gray in Figure 6 and are excluded from the scores presented in Table 3.

GPT-4o has a higher reported recall, precision, and F1 when the mean is computed across countries Table 3. Both models exhibit fairly low scores, indicating they miss a significant number of species. This difficulty seems to stem from the models' limitations when handling longer responses, which has been observed in other cases where LLMs struggle to generate or maintain coherence in extended outputs. In our case, the average response length was 2.3 species, with a maximum of 19 species identified across all countries. In regions like Indonesia, where up to 30 species need to be listed, this poses a particular challenge for both LLMs.

The models demonstrate high recall and F1 scores in several countries with few critically endangered (CR) species, as shown in Figure 6. Accurately predicting just a few CR species in these nations leads to strong performance based on our metrics. Many countries in southern Africa fall into this category. However, this performance is not uniformly high across all countries with low CR species counts and in many cases, such as Europe, F1 and recall are low. This inconsistency may arise because, when there are only a small number of correct species to predict, the model must be highly precise, making it challenging to generate outputs that align closely with the "ground truth". The models are able to achieve high F1 and recall in Brazil, India, and Vietnam, even though they are countries with a higher number of CR bird species.

TABLE 3 **Task 3: Threatened Species.** The mean precision, recall, and F1 computed across countries for `Gemini 1.5 Pro` and `GPT-4o`. Geographic regions and countries without any known CR birds are not included in these scores.

Model	Mean Precision	Mean Recall	Mean F1
`Gemini 1.5 Pro`	25.5%	18.0%	19.1%
`GPT-4o`	31.6%	20.1%	22.1%

3.4 | Task 4: Threat Assessment

Task 4 Prompt

```
Complete the threat assessment by making assessments for scope and severity for
the <common_name> (<scientific_name>).

Here are the scoring instructions from the IUCN: <scope_severity_instructions>

The threat description from IUCN is
<threats_description>.
Specifically, focus on the threat identified as <code> <title>. Additional details
about the threat code: \n{threat_code_description}

Scope options are:
- Whole (>90%)
- Majority (50-90%)
- Minority (<50%)

Severity options are:
- Very Rapid Declines
- Rapid Declines
- Slow, Significant Declines
- Causing/Could cause fluctuations
- Negligible declines
- No decline

Provide your answer as a flat Python list, containing only two elements like this:
[scope option, severity option]. Do not provide any additional context or
information.
```

FIGURE 7 **Prompt template for Task 4: Threat Assessment.**

In the context of species IUCN Red List assessments, detailed classifications of the threats each species faces are critical. While threat summaries are usually available, final threat assessments are often not provided, including classifications of each threat by "scope" (the proportion of the total population affected) and "severity" (the overall decline caused by the threat). For this task, we aim to assess an LLMs' ability to perform these scope and severity classifications.

Scope is the proportion of the population affected by the threat and is represented along with percentages, as this is how the information is provided by the IUCN's API. Scope can take on the values: "Whole (>90%)", "Majority (50-90%)", or "Minority (<50%)" (IUCN, 2024b). Severity is concerned with the acuteness of the threat, or the overall

TABLE 4 Task 4: Threat Classification. Accuracy (Acc.) and cost-sensitive accuracy (CS Acc.) for the scope and severity classification predictions made by `Gemini 1.5 Pro` and `GPT-4o`.

Model	Scope		Severity	
	Mean Acc.	Mean CS Acc.	Mean Acc.	Mean CS Acc.
`Gemini 1.5 Pro`	47.5%	65.2%	**28.3%**	**71.7%**
`GPT-4o`	**48.5%**	**66.7%**	22.2%	68.7%
Random Baseline	35.3%	57.1%	16.2%	51.8%

declines caused by it, and can take on the following categories: "Very Rapid Declines", "Rapid Declines", "Slow, Significant Declines", "Causing/Could Cause Fluctuations", "Negligible Declines", or "No Decline" (IUCN, 2024b). The species for this evaluation were chosen so that they covered a wide set of combinations of severity and scope classifications of threats, resulting in 100 severity and scope pairs. For instance, if hunting (i.e., threat code 5.1.1) is identified as a threat to a specific species, the model would be tasked with providing one of the defined scope options as well as one of the severity options relevant to that species. We selected these pairs independently of species names, meaning a single species may appear multiple times, each instance associated with different classifications of threat, severity, and scope. Thus we used these species with their known classifications to see whether the model, given the threat description, could replicate these classifications.

In the prompt, we provide the threat description for each species, along with an explanation of what the threat code means, including its descriptions and examples when available. Additionally, descriptions of what severity and scope entail are also included. The threat code descriptions are sourced from the IUCN Threat Classification documentation (IUCN, 2022).

In addition to evaluating the performance with accuracy reported separately for scope and severity, we implement a cost-sensitive performance metric to provide an approach that accounts for the proximity of errors for scope and severity classification. Instead of treating all incorrect predictions equally, we evaluate the degree of closeness between the predicted and true labels. For instance, in scope classification, predicting "Majority (50-90%)" when the true label is "Whole (>90%)" is considered less of an error compared to predicting "Minority (<50%)". Similarly, in severity classification, "Rapid Declines" is closer to "Very Rapid Declines" than to "Negligible Declines." This approach allows for partial credit when predictions are near the correct categories, offering a measure for the magnitude of errors, especially in cases where absolute accuracy may not reflect the model's ability to approximate the correct assessment. Exact implementation details are provided in Appendix B.

Neither model performs much better than a random guessing baseline on the task, as seen in Table 4. For the random guessing baseline, one scope and severity option is randomly selected for each threat assessment. For scope classification, `GPT-4o` outperforms `Gemini 1.5 Pro` with an accuracy of 48.5% compared to `Gemini 1.5 Pro`'s 47.5%. In severity classification, `Gemini 1.5 Pro` shows a small advantage, with higher accuracy (28.3%) and cost-sensitive accuracy (71.7%) compared to `GPT-4o`, but both barely outperform the random baseline. Although the cost-sensitive accuracy indicates that the models often misclassify into adjacent categories, the results suggest that relying solely on zero-shot capabilities based on threat descriptions is insufficient.

The models' performance, which is only about ten percentage points higher than the random baseline, suggests that LLMs struggle to reliably perform this task, which involves reasoning-like pattern recognition. However, the cost-sensitive accuracy, which is higher for both models than the random baseline, indicates that when the models are incorrect, they tend to be closer to the correct answer than random guessing, though not substantially better.

3.5 | Task 5: Trait Identification

Task 5 Prompt

```
Could you please provide an estimate of the <trait> (measured in <unit>) for the
species <common_name> (<scientific_name>)? Output your numerical estimate and
nothing else.
```

FIGURE 8 **Prompt template for Task 5: Trait Identification.**

This task is concerned with assessing LLMs' knowledge of traits of different species. To this end, we used the AVONET bird (Tobias et al., 2022) and COMBINE mammal (Soria et al., 2021) trait datasets. From AVONET, we extracted the beak, wing, and tail length, and mass for 100 different bird species. From COMBINE, we selected the adult body length, gestation length, max longevity, and adult mass for 100 mammal species. Neither of the datasets make any distinction between females and males for the traits, which has an impact on the expected values for some traits. The species were chosen for each dataset by randomly sampling 100 species. The two LLMs were queried with the prompt depicted in Figure 8.

We ask the models for a numeric answer so that we can easily parse it, but are lenient to responses that include the unit, e.g., `'50 mm'`. In those cases, we can retrieve the numeric part of the response. After querying the models we ended up with nine invalid responses out of 400 for `Gemini 1.5 Pro` across both bird and mammal tasks, and 0 for `GPT-4o`. The reasons why the model failed varied from including additional details, providing a range of values, or the model returning a null response.

For the bird traits, `GPT-4o` demonstrates an R^2 value closer to 1 and a lower Normalized Mean Absolute Error (NMAE) than `Gemini 1.5 Pro` across all traits, indicating that `GPT-4o` is more effective at predicting these traits. Out of the four traits, both models exhibit lower performance scores for beak length predictions. Additionally, `Gemini 1.5 Pro` struggles significantly with wing length predictions. Its erroneous forecast for the Turkey Vulture (*Cathartes aura*), which includes a measurement 1000 mm too large, results in a negative R^2 correlation score. For the mammal traits, `GPT-4o` outperforms `Gemini 1.5 Pro` across all four attributes. This includes the prediction of maximum longevity, which both models find particularly challenging. `GPT-4o` achieves an R^2 value of zero, while `Gemini 1.5 Pro` registers a negative R^2 score.

4 | DISCUSSION

The results of our study demonstrate both the promise and limitations of LLMs for providing knowledge and mimicking reasoning patterns in ecology. Here we discuss the performance of the models, next steps for developing more ecologically useful LLMs, and the implications for the ecological community.

Model performance. The performance of the tested versions of `GPT-4o` and `Gemini 1.5 Pro` demonstrated these models are very capable of successfully completing tasks such as species presence prediction and trait estimation, frequently exceeding simple baseline models. However, performance declined with increasing task complexity. This is highlighted by Task 2, the generation of range maps, where both `GPT-4o` and `Gemini 1.5 Pro` produced oversimplified or inaccurate outputs for most species. Comparing Task 2 with Task 1 not only highlights the notable performance contrast for both models under these different tasks, it also identifies a marked lack of internal consistency. Both models were capable of delivering reliable point location presence or absence predictions despite generating range

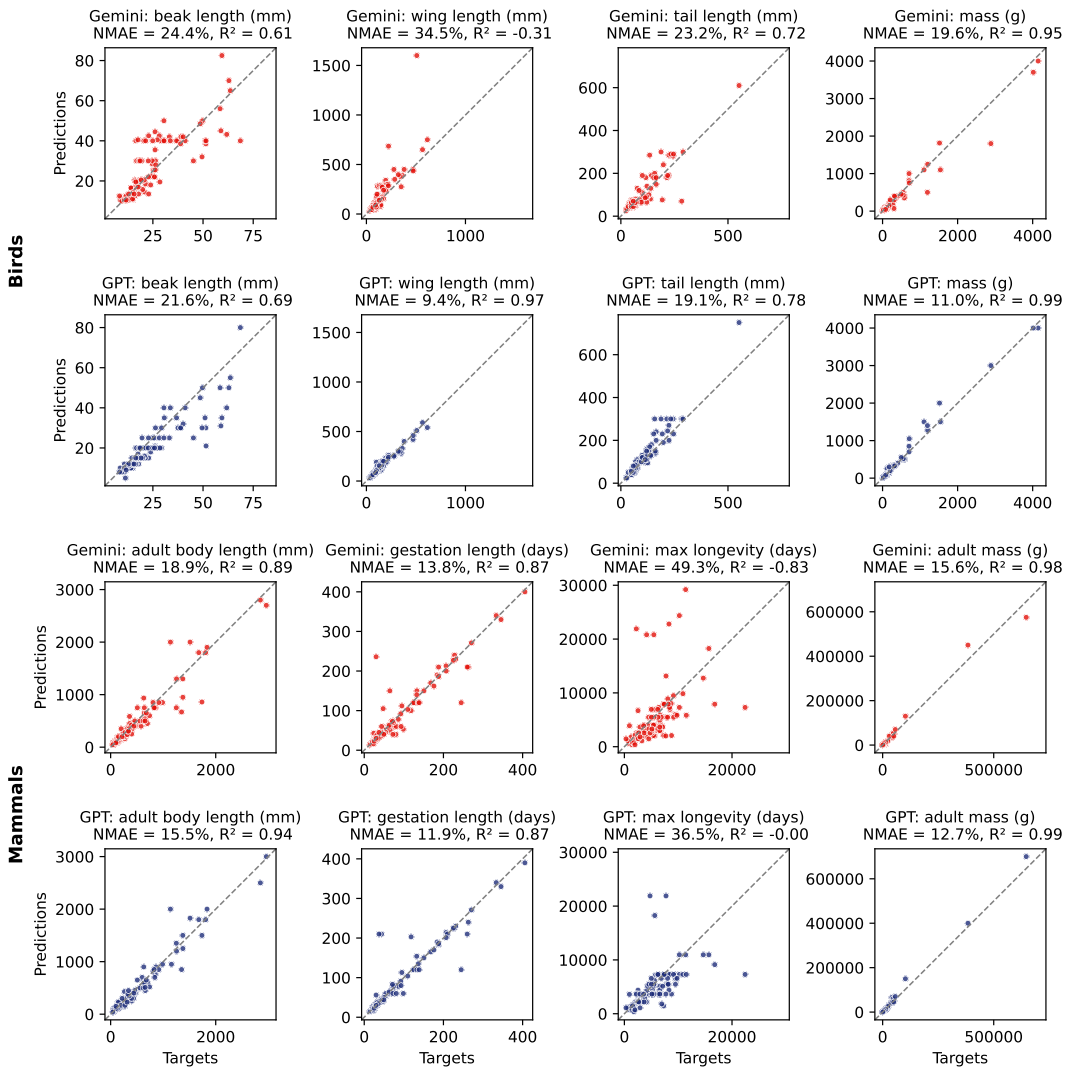

FIGURE 9 **Task 5: Trait Identification**. Comparison of predictions between the `Gemini 1.5 Pro` and `GPT-4o` models for four bird traits: beak length, wing length, tail length, and mass, as well as four traits for mammals: adult body length, gestation length, max longevity, and adult mass. Each panel shows a scatter plot of target values versus predictions with a reference line indicating a perfect prediction. NMAE (Normalized Mean Absolute Error) and R^2 scores are reported for both models.

maps with substantial flaws.

One explanation for poor performance, exhibited by more complex spatially grounded tasks is the nature of the underlying training data. LLMs are more likely to have encountered and retained information about countries and general species distributions in text, while specific geographic coordinates and highly detailed range maps are less prevalent in typical training data. The implications of the underlying training data may also be apparent in non-spatial

tasks. For instance, when estimating species traits, certain traits, like bird beak length, resulted in estimates that were clustered around a similar value across species, producing horizontal lines of points. This pattern was not observed for other traits, such as wing length, where the estimates varied more widely. These differences may reflect cases where the model has been trained on sufficient data to memorize the pattern (e.g., for wing length), as compared to cases where the model needs to infer the value (e.g., beak length). This is particularly interesting in the case of traits, and species locations, because there are known ecological relationships that the models could exploit. For example, there are general allometric relationships for beak length and wing length, resulting in a correlation between these two traits, which the model could leverage.

This gap between general knowledge and ecological or spatial reasoning underscores a core challenge in adapting LLMs for biodiversity conservation-related tasks. More contextual and domain-specific training may be necessary to improve performance, such as incorporating datasets specifically tailored to biodiversity threats and their attributes. This approach could also help address the limitations that both models exhibited when attempting to classify threats based on scope and severity. For example, in Task 4's scope and severity classification, it is worth noting that even experts may struggle to achieve perfect accuracy due to the complexity of the task. A more detailed analysis could delve into these inaccuracies, comparing model performance with human benchmarks and expert-level expectations.

This interplay of model limitations and task complexity is also evident in Task 3 (List Threatened Species), where performance metrics like recall and precision revealed their own strengths and shortcomings. A higher recall allowed `GPT-4o` to identify a larger proportion of relevant species, which is particularly valuable in conservation contexts where missing species could impact decision making. Accurate identification is critical for resource allocation and prioritization in conservation, as failing to identify key species risks missed opportunities for intervention or misguided actions. However, maximizing recall is not without its challenges, as it often comes at the cost of reduced precision. Achieving an optimal balance between recall and precision is crucial to minimize inefficiencies in resource allocation and ensure effective intervention strategies.

Ecological implications. There are several reasons why the development of an ecologically informed LLM could be useful. Biodiversity is a complex concept, and information about it is complicated, fragmented and often inaccessible. Lack of access to tools that can integrate, analyze and interpret biodiversity knowledge hinders better decisions. This is true even in developed nations and is further exacerbated in regions with fewer resources and diverse languages. A multi-model LLM that integrates current datasets, scientific reports, academic articles, and monitoring data into accurate and accessible biodiversity intelligence would be a hugely valuable support tool for decision makers across organization types.

On the other hand, LLMs raise several concerns. There are important ethical issues, for example around amplifying biases and generating misleading information (Tokayev, 2023). In the ecological domain, there is also a concern about the appropriation of local and indigenous knowledge. More broadly, LLMs have large energy consumption and associated environmental footprints, both during training and inference (Samsi et al., 2023). This makes decisions about when to use LLMs for ecological applications important. For instance, if aiming to look up critically endangered birds of Brazil, a simple web search might return the required information at a much lower energy consumption compared to asking an LLM.

In our study, we utilize closed commercial LLMs. That means that the implementation details and data they have been trained on are not publicly available. This choice is motivated by the fact that we wish to evaluate LLMs that represent the best performance currently available. We make our code and evaluation data openly available to allow others to extend and build upon our analysis, e.g., by applying it to other models.

Developing more ecologically useful LLMs. Our evaluation demonstrates the already impressive capabilities of current LLMs for ecological applications. The development of more ecologically informative LLMs offers significant benefits

for research and conservation but also raises ethical and environmental concerns. As their usefulness for ecological purposes grows, it is crucial to balance these benefits with responsible development. Geospatially explicit AI models have recently emerged and it is likely these will be incorporated into multi-modal LLMs, enabling better performance in geospatial tasks. So, it seems prudent to ask what the next steps need to be to develop ecologically useful LLMs whilst minimizing some of the ethical and environmental risks.

In the near term, the analysis we present here could be advanced by evaluating how changes in prompt structure affect the performance of these models. Likewise, the analysis could also be extended to a few-shot setting whereby examples are provided together with the prompts. For the longer-term endeavor of creating ecologically useful LLMs, an important development will be training models to understand the geographic, taxonomic, temporal, and knowledge systems gaps and biases in current biodiversity knowledge. This will be critically important to avoid amplifying the biased knowledge contained in the training data. Part of this development could be achieved by learning from the ecological theory and documentation of the sampling biases that exist. To combat the generation of misleading information from future models, a formal set of evaluation benchmarks is needed, building upon the evaluation that we have introduced here.

These benchmarks require a careful formulation to span a range of tasks that an LLM might be required to perform. It is also important to take into account limitations in current data and to evolve as the state of knowledge changes. For example, the expert-derived data we use for evaluation (e.g., the range maps), though representative of the currently best available knowledge, are not guaranteed to be fully accurate. Species threat assessments are dynamic and should respond to changing natural and human pressures on a species. Therefore the assessments we used here as the "ground truth" are not guaranteed to reflect the current status of a species. However, the data used is some of the best available for evaluating model outputs.

Our evaluation focuses mostly on retrieval-style tasks where the information required to answer the question could conceivably be found on the internet. It is not inconceivable that the LLMs we evaluated have encountered such data during training, which underscores an important aspect of these models: their ability to leverage previously learned information to answer questions. However, this also raises the concern that the models may be merely reproducing patterns they have seen, rather than extrapolating from this information to make reasoning-like predictions. A more comprehensive evaluation in the future should explore tasks that require more complex ecological reasoning to perform. For example, predicting outcomes of positive or negative impacts on ecosystems, imputing ecological information taking into account phylogenetic, functional, and geographic knowledge. However, the challenge with such tasks is that it is not necessarily clear how to evaluate them in a scalable and reproducible manner.

The evaluation and comparison of LLMs on a fixed dataset is a necessary and objective measure of an LLMs' ability to accurately recall ecological information. However, another aspect of evaluation that is important to consider is that the current alternative is using experts to synthesize information. An interesting evaluation of an ecologically useful LLM would be to compare its ability against human experts. In many cases, we expect that even the zero-shot LLMs that we evaluate here are performing more accurately across a range of tasks, than an expert could.

5 | CONCLUSIONS

We have presented an initial evaluation of LLMs in the ecological domain, which highlights both their current abilities, current limitations, and points to their potential in the future. The commercial models evaluated show promise for tasks that involve recalling information, such as species presence predictions and trait estimation. However, their performance drops significantly in more complex tasks, such as generating accurate range maps or detailed threat

classifications.

While these models could assist in accelerating certain conservation-related tasks and already make ecological information more accessible, their limitations suggest that they are not yet ready for direct application in biodiversity decision making. Nonetheless, our new evaluation dataset represents a valuable resource that can be used to stimulate and facilitate future research, enabling advancements in the application of LLMs to ecological and conservation challenges. Models that are more ecologically useful are likely to improve over time as techniques and training datasets evolve. However, it is important that we direct development towards ethically sound practices. Given the urgency of the biodiversity crisis, the benchmarking tasks introduced here can catalyze efficient and focused development of tools to help accelerate better conservation decision making.

Conflict of Interest

The authors declare that they have no conflicts of interest related to this work.

Author Contributions

FD, JM, DP, MH, OMA together developed the concept and designed the methodology. FD implemented the methodology and carried out the experiments and analysis. JM, DP, MH, OMA supervised the study. All authors contributed to writing the manuscript based on the initial draft by FD and gave final approval for publication

Data Availability

The benchmark dataset and code are publicly available at https://github.com/filipgdorm/eco-llm. The README in the linked page provides installation instructions and further details about the evaluation data and how it was collected.

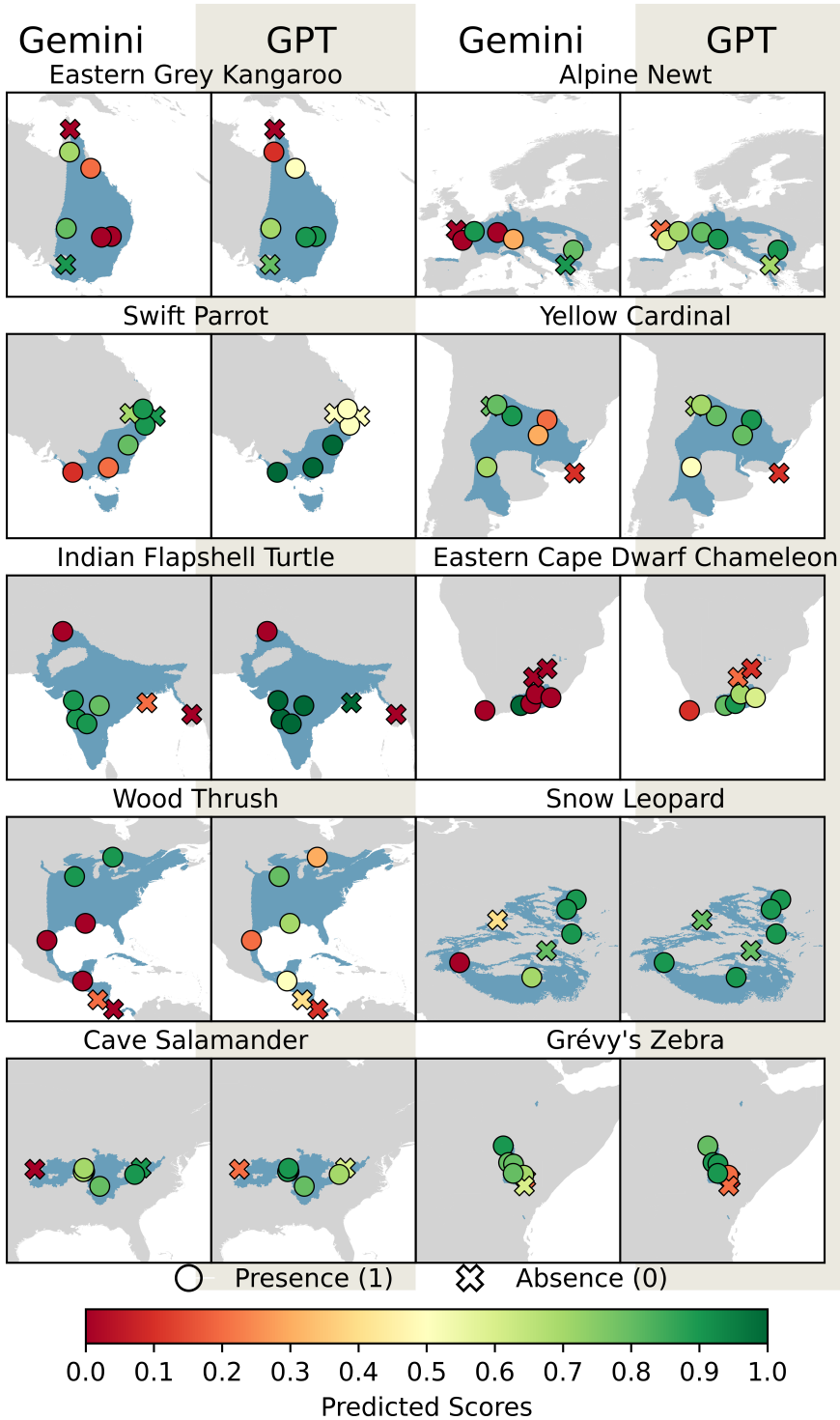

FIGURE 10 **Task 1: Species Presence**. Visualization of the locations and correct/incorrect classifications for species on the map. Shape corresponds to presence/absence, color code the predicted score from model. The map displays presences and only the challenging nearby absences.

References

Atuhurra, J., Dujohn, S. C., Kamigaito, H., Shindo, H. and Watanabe, T. (2024) Distilling named entity recognition models for endangered species from large language models. *arXiv:2403.15430*. 2

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C. et al. (2011) Has the earth's sixth mass extinction already arrived? *Nature*. 1

Castro, A., Pinto, J., Reino, L., Pipek, P. and Capinha, C. (2024) Large language models overcome the challenges of unstructured text data in ecology. *bioRxiv*. 2

Chang, K. K., Cramer, M., Soni, S. and Bamman, D. (2023) Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv:2305.00118*. 2

ChatGPT (2024) https://chat.openai.com. 3

Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S. and Mac Aodha, O. (2023) Spatial implicit neural representations for global-scale species mapping. In *ICML*. 3, 6

Cooper, N., Clark, A. T., Lecomte, N., Qiao, H. and Ellison, A. M. (2024) Harnessing large language models for coding, teaching and inclusion to empower research in ecology and evolution. *Methods in Ecology and Evolution*. 2

Deng, C., Zhang, T., He, Z., Chen, Q., Shi, Y., Xu, Y., Fu, L., Zhang, W., Wang, X., Zhou, C. et al. (2024) K2: A foundation language model for geoscience knowledge understanding and utilization. In *International Conference on Web Search and Data Mining*. 2

Doi, H., Osawa, T. and Tsutsumida, N. (2024) The role of large pre-trained models in ecology and biodiversity conservation: Opportunities and challenges. *Authorea Preprints*. 2

Domazetoski, V. (2024) Enhancing ecological knowledge discovery using large language models. *MSc Thesis*. 2

Dorm, F., Lange, C., Loarie, S. and Mac Aodha, O. (2024) Generating Binary Species Range Maps. In *Computer Vision for Ecology Workshop at ECCV*. 3, 6, 7

Fabian, Z., Miao, Z., Li, C., Zhang, Y., Liu, Z., Hernández, A., Montes-Rojas, A., Escucha, R., Siabatto, L., Link, A. et al. (2023) Multimodal foundation models for zero-shot animal species recognition in camera trap images. *arXiv:2311.01064*. 2

Farrell, M. J., Le Guillarme, N., Brierley, L., Hunter, B., Scheepens, D., Willoughby, A., Yates, A. and Mideo, N. (2024) The changing landscape of text mining: a review of approaches for ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*. 2

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. and Berner, J. (2023) Mathematical capabilities of chatgpt. *NeurIPS Datasets and Benchmarks*. 2

Gemini (2024) https://gemini.google.com. 3

Gemini - API (2024) https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/inference. 3

Gougherty, A. V. and Clipp, H. L. (2024) Testing the reliability of an ai-based large language model to extract ecological information from the scientific literature. *npj Biodiversity*. 2

GPT-4 - API (2024) https://platform.openai.com/docs/overview. 3

H3 (2024) https://h3geo.org. 4

Hamilton, M., Lange, C., Cole, E., Shepard, A., Heinrich, S., Mac Aodha, O., Van Horn, G. and Maji, S. (2024) Combining observational data and language for species range estimation. In *NeurIPS*. 2

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M. and Ladle, R. J. (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual review of ecology, evolution, and systematics*. 1

IUCN (2022) Iucn - cmp unified classification of direct threats. https://nc.iucnredlist.org/redlist/content/attachment_files/Dec_2022_Guidance_Threats_Classification_Scheme.pdf. 10

— (2024a) The iucn red list of threatened species. URL: https://www.iucnredlist.org. Version 2024-1. 3

— (2024b) Threats classification scheme (version 3.3). https://www.iucnredlist.org/resources/threat-classification-scheme. 9, 10

IUCN Red List - API (2024) Https://apiv3.iucnredlist.org/api/v3/docs. 3

Le Guillarme, N. and Thuiller, W. (2022) Taxonerd: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*. 3

Manvi, R., Khanna, S., Burke, M., Lobell, D. and Ermon, S. (2024) Large language models are geographically biased. *arXiv:2402.02680*. 2

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D. and Ermon, S. (2023) Geollm: Extracting geospatial knowledge from large language models. *arXiv:2310.06213*. 2

Marcos, D., van de Vlasakker, R., Athanasiadis, I. N., Bonnet, P., Goeau, H., Joly, A., Kissling, W. D., Leblanc, C., van Proosdij, A. S. and Panousis, K. P. (2024) Fully automatic extraction of morphological traits from the web: utopia or reality? *arXiv preprint arXiv:2409.17179*. 2

Maruf, M., Daw, A., Mehrab, K. S., Manogaran, H. B., Neog, A., Sawhney, M., Khurana, M., Balhoff, J. P., Bakis, Y., Altintas, B. et al. (2024) Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. *arXiv:2408.16176*. 2

Miao, Z., Zhang, Y., Fabian, Z., Celis, A. H., Beery, S., Li, C., Liu, Z., Gupta, A., Nasir, M., Li, W. et al. (2024) New frontiers in ai for biodiversity research and conservation with multimodal language models. *EcoEvoRxiv*. 2

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S. and Farajtabar, M. (2024) Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv:2410.05229*. 2

Mora-Cross, M. and Calderon-Ramirez, S. (2024) Uncertainty estimation in large language models to support biodiversity conservation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2

Poldrack, R. A., Lu, T. and Beguš, G. (2023) Ai-assisted coding: Experiments with gpt-4. *arXiv:2304.13187*. 2

Roberts, J., Lüddecke, T., Das, S., Han, K. and Albanie, S. (2024a) Gpt4geo: How a language model sees the world's geography. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*. 2

Roberts, J., Lüddecke, T., Sheikh, R., Han, K. and Albanie, S. (2024b) Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *CVPR*. 2

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D. and Gadepally, V. (2023) From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–9. IEEE. 13

Santangeli, A., Mammola, S., Nanni, V. and Lambertucci, S. A. (2024) Large language models debunk fake and sensational wildlife news. *Integrative Conservation*. 2

Sastry, S., Xing, X., Dhakal, A., Khanal, S., Ahmad, A. and Jacobs, N. (2023) Ld-sdm: Language-driven hierarchical species distribution modeling. *arXiv:2312.08334*. 2

Scheepens, D., Millard, J., Farrell, M. and Newbold, T. (2024) Large language models help facilitate the automated synthesis of information on potential pest controllers. *Methods in Ecology and Evolution*. 2

Soria, C. D., Pacifici, M., Di Marco, M., Stephen, S. M. and Rondinini, C. (2021) Combine: a coalesced mammal database of intrinsic and extrinsic traits. *Ecology*. 3, 11

Tao, R. and Xu, J. (2023) Mapping with chatgpt. *ISPRS International Journal of Geo-Information*. 2

Tobias, J. A., Sheard, C., Pigot, A. L., Devenish, A. J., Yang, J., Sayol, F., Neate-Clegg, M. H., Alioravainen, N., Weeks, T. L., Barber, R. A. et al. (2022) Avonet: morphological, ecological and geographical data for all birds. *Ecology Letters*. 3, 11

Tokayev, K.-J. (2023) Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*, **8**, 17–33. 13

Zhang, Y., Wang, Z., He, Z., Li, J., Mai, G., Lin, J., Wei, C. and Yu, W. (2024) Bb-geogpt: A framework for learning a large language model for geographic information science. *Information Processing and Management*. 2

Supporting Information

A | DATASETS

The full list of species utilized for the tasks is available in the `SPECIES.md` file within the code repository.

B | EVALUATION METRICS

F1 Score. The F1 score is computed in the overlap of cells between the LLM generated range map, and the expert range map.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}, \tag{1}$$

where TP denotes True Positives, FP False Positives, and FN False Negatives.

Cost Sensitive Accuracy Calculation. In Task 4: Threat Classification, instead of treating all incorrect predictions equally, we evaluate the degree of closeness between the predicted and true labels. For instance, in scope classification, predicting "Majority (50-90%)" when the true label is "Whole (>90%)" is considered less of an error compared to predicting "Minority (<50%)." Similarly, in severity classification, predicting "Rapid Declines" is closer to "Very Rapid Declines" than to "Negligible Declines."

To quantify this proximity, we utilize a scoring system that assigns weights to the categories based on their relative relationships. We define a similarity function for both scope and severity, where exact matches receive a full score, and adjacent or nearby categories are awarded partial credit.

For scope classification, a mapping is created where "Whole" is assigned a score of 3, "Majority" a score of 2, and "Minority" a score of 1. If the predicted label matches the correct label, a score of 1.0 is given for an exact match. If the predicted label is adjacent to the correct label, a score of 0.5 is assigned. If the categories are far apart, a score of 0 is given.

For severity classification, a similar mapping is established with "Very Rapid Declines" receiving a score of 6, "Rapid Declines" a score of 5, "Slow, Significant Declines" a score of 4, "Causing/Could Cause Fluctuations" a score of 3, "Negligible Declines" a score of 2, and "No Decline" a score of 1. Exact matches receive a score of 1.0. If the predicted label is close to the correct label, it receives a score of 0.75; for intermediate matches, a score of 0.5 is awarded. A score of 0.25 is assigned for distant matches, and a score of 0 is given when the categories are at opposite ends.

This approach allows for partial credit when predictions are near the correct categories, offering a measure for the magnitude of errors. This is particularly useful in scenarios where absolute accuracy may not fully capture the model's performance in approximating the correct assessment.