Large language models overcome the challenges of unstructured text data in ecology

Andry Castro[1]; João Pinto[2]; Luís Reino[3, 4, 5]; Pavel Pipek[6,7] César Capinha[1,8*]

[1]Centro de Estudos Geográficos, Instituto de Geografia e Ordenamento do Território, Universidade de Lisboa, Lisboa, Portugal.

[2]Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisboa, Portugal.

[3]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal.

[4]BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Vairão, Portugal.

[5]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa, Portugal.

[6]Department of Invasion Ecology, Institute of Botany, Czech Academy of Sciences, Průhonice, Czech Republic.

[7]Department of Ecology, Faculty of Science, Charles University, Prague, Czech Republic.

[8]Associated Laboratory Terra, Portugal.

[*]Corresponding author (cesarcapinha@campus.ul.pt)

ABSTRACT

The vast volume of unstructured textual data, such as that found in research papers, news outlets, and technical reports, holds largely untapped potential for ecological research. However, the labour-intensive nature of manually processing such data presents a considerable challenge. In this work, we explore the application of three state-of-the-art Large Language Models (LLMs) — ChatGPT 3.5, ChatGPT 4, and LLaMA-2-70B — to automate the identification, interpretation, extraction, and structuring of relevant ecological information from unstructured textual sources. Our focus is specifically on species distribution data, using two challenging sources of these data: news outlets and research papers. We assess the LLMs on four key parameters: identification of documents providing species distribution data, identification of regions where species observations are mentioned, generation of geographical coordinates for these regions, and provisioning of results in a structured format. Our results show that ChatGPT 4 consistently outperforms the other models, demonstrating a high capacity to interpret textual narratives and to extract relevant information, with a percentage of correct outputs often exceeding 90%. However, performance also seems dependent on the type of data source used and task tested – with better results being achieved for news texts and in identifying regions where species were observed and presenting structured output. Its predecessor, ChatGPT 3.5, delivers reasonably lower accuracy levels across tasks and data sources, while LLaMA-2-70B performed worse. The integration of LLMs into ecological data assimilation workflows appears not only imminent, but also essential to meet the growing challenge of efficiently processing an increasing volume of textual data.

KEYWORDS: AI; automation; data integration; GPT; LLaMA; unstructured data;

1 | INTRODUCTION

In the current Information Age, we are experiencing an unprecedented increase in the volume and diversity of data made available by an also increasing number and diversity of sources (Hampton et al., 2013). This wealth of data holds a large potential for scientific research, offering the opportunity for significant breakthroughs across research domains (Grossi et al., 2021). However, the unstructured nature of much of the data being delivered presents significant analytical challenges.

Ecological research offers many examples of the added value of the large volume of unstructured data now available. Data is considered 'unstructured' if it is not arranged according to the research analytical categories (Boulton & Hammersley, 2006). In ecology this includes a vast array of data types that play a central role in research, such as remote sensing or citizen science data (Bayraktarov et al., 2019), or for which data integration and analytical pipelines are advancing at a fast pace, such as acoustic data (Sethi et al., 2020). However, other types of unstructured data are lagging significantly behind, despite their enormous research potential. This is the case of textual data, as provided in research papers, news outlets or technical reports, and consisting mainly of free-flowing text or text in tables with multiple layouts. Previous research has identified the significance of these data for ecological research and applications including sentiment analysis, text mining and species distribution mapping (e.g., Hart et al., 2018; Moloney et al., 2021; Monteiro et al., 2020). However, despite some noteworthy recent improvements of available tools, for example allowing to identify mentions of specific taxa in text (e.g., 'Taxonerd'; Le Guillarme & Thuiller, 2022), or identifying research articles relevant for ecological data assembly (Cornford et al., 2020), the work (going

from information identification, extraction, and harmonisation) still requires extensive human input.

A particular example of the challenge posed by unstructured textual data in ecology concerns species distribution data, which is crucial for many primary applied and basic research questions, such as understanding biodiversity loss and change (e.g., Boonman et al., 2024) and in assessing the spread of invasive alien species (IAS) (Latombe et al., 2017). Currently, we are experiencing a flood of unstructured textual data informing about species distributions, coming from sources like the social media (Chowdhury et al., 2023), research papers (e.g., Maquart et al., 2021), technical reports (e.g., Mota et al., 2006) and webpages of institutions (e.g., gencat.cat). However, much of these data is either being ignored by researchers (opting, for example, to use only readily-usable structured observation data, as provided by GBIF; gbif.org), or its assimilation and use is made with a high temporal latency, owing to the dependence on manual, labour-intensive procedures for identification, extraction and harmonisation with structured data. For a concrete example, our team has developed several global-scale datasets of the distribution of non-native taxa (e.g., Capinha et al., 2017, 2020; Monteiro et al., 2020) for which we had to spend several hundreds of hours to identify, assimilate and integrate the relevant information that was provided in the form of unstructured text (e.g., species distribution narratives in research papers or monographs, or tables indicating species occurrence or absence in multiple regions). Worryingly, the rapidly growing volume of scientific and non-scientific publications (Landhuis, 2016) and as well as a low adoption of data reporting standards for the former (Castro et al., 2023; Poisot et al., 2019) is likely to further exacerbate this challenge as time progresses.

In this context, the recent development and public availability of advanced large language models (LLMs), such as GPT, offers a promising solution. These models have demonstrated a revolutionary ability to retrieve and analyse natural language, including its context and distinct nuances in meaning (Kheiri & Karimi, 2023). Trained on enormous corpora of text, including scientific literature, and other data, the models also have a proven ability to analyse and transform the data, potentially delivering it in structured forms (Brown et al., 2020; OpenAI, 2023; Radford et al., 2018, 2019). These capabilities hold the promise of LLMs as a comprehensive tool for data identification, extraction, and structuring from textual data in ecology. However, it is also apparent that their performance in these tasks depends on the models used, the type of data, and the desired structure of the extracted information (Mao et al., 2023). Therefore, the time is opportune to evaluate the capacity of these models for automated textual data assimilation and structuring in ecological research.

Our work aims to evaluate the capability of large language models (LLMs) in identifying, extracting, and structuring relevant ecological information from unstructured textual data. Specifically, we compare the performance of three state-of-the-art LLMs - ChatGPT 3.5, ChatGPT 4 (OpenAI, 2023), and LLaMA-2-70B (Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023) - in their ability to identify, extract, georeferenced and formatted species distribution data from free-flowing text.

2 | METHODOLOGY

To achieve our goals, we simulate the application of LLMs in a real-world scenario, employing them to process unstructured textual data in news outlets and research papers about two range-expanding, problematic invasive species. The goal is to extract

structured distribution data from these sources to better track the species ongoing range expansion. We evaluate four key parameters, covering the spectrum from information identification to integration: the ability of models to 1) identify research papers or news articles in different media outlets that provide species observation data, 2) extract the name of the regions where the observations were made, 3) generate geographical coordinates for these regions, and 4) present the information in a structured, consistent format.

2.1 | Building a testing dataset

We used the Asian hornet (*Vespa velutina*) and the Asian tiger mosquito (*Aedes albopictus*) as test cases for our work. These species are highly problematic in their alien range, causing economic damage and threat to human health (Barbet-Massin et al., 2020; Schaffner et al., 2013), and they are also spreading at a fast rate to new regions, particularly in Europe (Monceau & Thiery, 2016; Schaffner & Mathis, 2014). Because of this, there is significant interest of researchers and entities tasked with invasion prevention strategies, in being able to track the geographical progression of these species. However, a relevant volume of distributional information being provided comes in the form of textual narratives delivered in news (e.g., Bullens, 2023; Carballo, 2023) or in research papers (e.g., Bakran-Lebl et al., 2021; Dillane et al., 2022).

To mimic the current situation of unstructured textual data being provided on the distribution of these species, we considered two types of sources: research papers and online newspaper news. Research papers publish a relevant number of works for these species, however only a limited number of them tend to provide distributional data, with other common focus being health or economic impact-related information, genetic

studies, etc. Similarly, online newspaper news also provides key and often timely information on new species observations (e.g., Vuorisalo et al., 2001). However, this is often mixed with news on several other subjects, even if same-species related, such as precautions to be taken and potential economic or health impacts.

To incorporate these two types of data source, we used two programmatic interfaces for the R Language (R Core Team, 2022): the 'openalexR' (Aria et al., 2023) and the 'newsanchor' (Frie et al., 2019) packages. The first, is an R wrapper for the Open Alex API (https://openalex.org) which enables users to query and extract data from the Open Alex database which includes over 250M academic works. It provides access to multiple fields of these works, including titles, abstracts, publication dates, and more. The 'newsanchor' package connects to the News API (https://newsapi.org), allowing it to search and retrieve live news headlines from over 30,000 news sources and blogs. Here we used the developer licence of the News API, which includes some limitations, including searching only news up to a month old. Crucially, both packages allow near-real-time access to published information, making them suitable for continuous and timely monitoring of new species distribution data.

We searched for research papers and news using scientific name and common names of species in English in case of research papers and in ten different languages for the news (Dutch, English, French, Italian, German, Norwegian, Portuguese, Russian, Spanish and Swedish; e.g., *Aedes albopictus*, tiger mosquito, zanzara tigre, asiatisk tigermygg). For research papers we saved titles, abstracts and DOIs, while for news we saved its title and description ('the highlight'). Due to limitations in API access and unequal volumes of results, the period used in the searches differed between species and source type. In each case, we extended the temporal extent preceding the date of search retrieval until a

total of 150 results were obtained. This resulted in a data set of 600 records to test (2 × 150, for each species; **Table 1**). For each of these records, we then manually identified if it reported species' distributional information, and if so, for which region or regions (**Table 1** and see the full database of news and papers in the **Appendix 1**).

Table 1 – Composition of the data set assembled to test the capacity of large language models in identifying distribution data and information extraction for two invasive species: the *Aedes albopictus* mosquito and the hornet *Vespa velutina*.

Species	Data source	Reporting of distribution data		
		Yes	No	Total
Aedes albopictus	News	44	106	150
	Research papers	61	89	150
Vespa velutina	News	79	71	150
	Research papers	65	85	150

2.2 | Prompt approach

We implemented a few-shot learning approach (Brown et al., 2020; Chiu et al., 2021), giving LLMs examples of data sources reporting or not species distributions. We performed this separately for research papers and for news. For the former, we provided a set of eight positive and eight negative cases, with each characterised by title and abstract. For news, we used the same procedure, giving the news title (in some cases only, as this field is not always provided by the search results), alongside the description ('highlight') of the news.

Then, we instructed the models to generate a simple answer of "YES" or "NO" to identify if papers or news to be examined include or not (respectively) distribution data of the species. Then, for positive classifications, we also instructed the model to provide

the name or names of regions where the species is mentioned to occur and, in the following line, to provide its geographical coordinates in decimal degrees. We also provide the models with examples of the expected result in both cases. Additional species-relevant instructions were also given. Specifically, we informed about possible confusions between species common names (e.g., 'Asian hornet' vs 'Asian giant hornet', the latter being a distinct species) and to exclude research papers related to potential (i.e., not actually observed) distributions and papers or news referring to geographical resolutions coarser than that of a single country (e.g., a set of two or more countries, or a continent). Finally, we also asked the model to provide, the results in a consistent, standardised, structure, having the classification result (i.e., 'YES' or 'NO'), in the first line, and for positive cases the region name(s) in the second line and the coordinates in the third (last) line (See the full text prompt in the **Appendix 2**).

2.3 | Data testing

We performed our tests using three recent large language models: GPT-3.5 and GPT-4 from OpenAI (OpenAI, 2023), and LLaMA-2-70B (Touvron, Martin, et al., 2023) from Meta AI. The Generative Pre-trained Transformer (GPT) series, based on the Transformer architecture introduced by Vaswani et al. (2017), represents a major advancement in the field of LLMs. The GPT-3 model, presented by Brown et al. (2020), features an autoregressive LM with 175 billion parameters, trained on a vast text corpus. Its enhanced version, GPT-3.5, incorporates reinforcement learning from human feedback (Ouyang et al., 2022) to improve performance and is currently available in a limited chat mode from OpenAI. The subsequent model, GPT-4, is larger and more advanced, though its exact parameter count remains undisclosed. The use of GPT-4 is restricted to commercial licences, and it is considered the current state-of-the-art in

large LLMs. LLaMA-2-70B, an open-source LLM, is pretrained on 2 trillion tokens of data, including instruction datasets and human-annotated examples. This model outperforms other open-source LLM models on most benchmarks (Touvron, Martin, et al., 2023), and aims to be a suitable substitute for closed-source models such as those from the GPT family. The first two (GPT) models were tested through chat.openai.com interface and llama2.ai interface was used for LLaMA-2-70B tests.

For each model, we performed 600 individual tests, each in a new dialog window, to guarantee that the answer was not influenced by previous ones. For each we evaluated 1) the correct classification of the news or paper in terms of the provision or not of distribution data 2) the correct naming of regions for which distribution records are reported, 3) the provision of geographical coordinates falling within the region for which the distribution is reported and 4) the provision of results in the specified structured format.

We assessed models' performance by measuring the accuracy (i.e., percentage of correct results) and complementary true positive and true negative percentages, for each of the evaluated parameters. For geographical coordinates, we also compared the performances obtained with those obtained for the current state-of-the-art approach, consisting of geocoding regions identified by the LLMs using Nominatim 4.3.2 API (https://nominatim.org/release-docs/latest/api/), an open-source geocoding service provided by the OpenStreetMap (OSM) project, via R.

Examples of how the results were evaluated are shown in **Table 2**. To ease interpretation and communication of performances, we adopted a qualitative

classification of accuracy values, corresponding to perfect (100%), very good (<100% and ≥ 90%), good (<90% and ≥ 70%), moderate (<70% and ≥ 50%) and poor (<50%).

Table 2 – Examples of the evaluation of results obtained from the models, referring to the classification of documents, the correct identification of regions for which the species distribution is mentioned, the provision of coordinates within the mentioned region and the provision of all results in the requested data arrangement structure. Two cases of news reporting the presence of *Vespa velutina* in Jersey Island and a single example of news about *Aedes albopictus* are provided. The first two correspond to positive cases (i.e., species distribution is reported) and the latter to a negative case, for which it is not possible to establish a relationship between the species and a specific geographical area.

Examples to classify	**Expected result**	**Parameters classified**			
		Document classification	Region name	Coordinates	Structure requested
Example of an expected "YES", directly reported					
"Invasive hornet species found in the Jersey Island for the first time - This marks the first time the yellow-legged hornets have been reported on the Island."	"YES Jersey Island 49.214439°, -2.131250°"	Correct	Correct	Correct	Correct
Example of an expected "YES", indirectly reported					
"Horrifying moment Asian hornet devours wasp in a Jersey Island school in just seconds -The video shows the moment that an Asian hornet devoured a wasp in seconds"	"YES Jersey Island 49.214439°, -2.131250°"	Correct	Correct	Correct	Correct
Example of an expected "NO"					
"Aedes albopictus is growing in urban areas of developed countries - Cities are facing a threat under a new climatic conditions"	"NO"	Correct	NA	NA	Correct

3 | RESULTS

The tests performed show that GPT-4 typically outperforms GPT-3.5, and both consistently surpass LLaMA-2-70B across the four assessed parameters **(Figure 1)**.

3.1 | Classification of documents

The first parameter assessed was the capacity of models to identify unstructured data sources which are reporting the presence of a given species **(Figure 1a)**. GPT-4 presents good and very good levels of accuracy for news (88% for news about *Aedes albopictus* and 92.7% for news related with *Vespa velutina*) and good levels of accuracy for papers (83.3% in papers about *Aedes albopictus* and 82% for papers related with *Vespa velutina*). GPT-4 is followed by GPT-3.5 with very good levels of accuracy for news (*Aedes albopictus* = 92.7% and *Vespa velutina* = 90%) and good levels for papers (*Aedes albopictus* = 74% and *Vespa velutina* = 70.7%). The LLaMA-2-70B had the least accurate results for this parameter, with good levels for news (*Aedes albopictus* = 78.7% and *Vespa velutina* = 72%) and poor to moderate levels for papers (*Aedes albopictus* = 48% and *Vespa velutina* = 55.3%). Notably, GPT-4 was more accurate at identify documents that do not report species presence (94.7%± 4.6) compared to those that do (73.4%±16.6). In contrast, GPT-3.5 showed a smaller average difference (81.9%±12.4 for true negatives vs. 79.9%±19.9 for true positives). The LLaMA-2-70B model showed the opposite trend, being more accurate in identifying true positives (69.9%±12.0) than true negatives (54.4%±32.1), though its overall performance was markedly lower **(Table 3)**.

3.2 | Geographical features

The second and third parameters assessed referred only to news or papers reporting the species considered. Thus, considering the naming of the region **(Figura 1b)**, GPT-4 also leads with perfect levels of accuracy for news (*Aedes albopictus* = 100% and *Vespa velutina* = 100%) and papers (*Aedes albopictus* = 100% and *Vespa velutina* = 100%). GPT-3.5 achieved very good to perfect levels of accuracy for news (*Aedes albopictus* = 97.5% and *Vespa velutina* = 100%) and very good levels for papers (*Aedes albopictus* = 96.9% and *Vespa velutina* = 92.2%). Once more, LLaMA-2-70B was the weakest model, but still with very good and good levels of accuracy for news (*Aedes albopictus* = 95.2% and *Vespa velutina* = 84.8%), but moderate and poor levels for papers (*Aedes albopictus* = 61.7% and *Vespa velutina* = 49.1%).

Considering now the capacity to provide coordinates falling within the regions identified **(Figure 1c)**, GPT-4 also leads with perfect and very good levels of accuracy for news (*Aedes albopictus* = 100% and *Vespa velutina* = 96.1%) and good and very good levels of accuracy for papers (*Aedes albopictus* = 89.2% and *Vespa velutina* = 95%). GPT-3.5 reached very good levels of accuracy for news (*Aedes albopictus* = 92.5% and *Vespa velutina* = 92.2%), and good levels for papers (*Aedes albopictus* = 87.5% and *Vespa velutina* = 88.2%). Finally, LLaMA-2-70B presented the lowest accuracy levels for this parameter, reaching good and moderate values for news (*Aedes albopictus* = 84.6% and *Vespa velutina* = 56.5%), but poor levels for papers (*Aedes albopictus* = 36.2% and *Vespa velutina* = 43.4%).

The results of GPT-4 are generally comparable to those achieved by the current state-of-the-art geocoding method that involves the use of Nominatim. Considering all positive cases in a structured format, which GPT-4 assessed correctly and coordinates (n=177),

and positive ones that were well-formatted with a correct region but wrong coordinates (n=9), that the accuracy of GPT-4 (95.2%) is similar to the accuracy of Nominatim (96.2%).

3.3 | Provision of results in a structured format

Concerning the provision of results in the structure requested **(Figure 1d)**, we verified 100% of accuracy for both data source types using GPT-4. GPT-3.5 achieved very good and very good levels of accuracy in news (*Aedes albopictus* = 94% and *Vespa velutina* = 92.7%) and good to moderate levels in papers (*Aedes albopictus* = 72.7% and *Vespa velutina* = 64%). LLaMA 2-70B model is again the least performing, reaching poor levels of accuracy in news (*Aedes albopictus* = 48% and *Vespa velutina* = 37.3%) and poor levels in papers (*Aedes albopictus* = 26.7% and *Vespa velutina* = 34.7%).

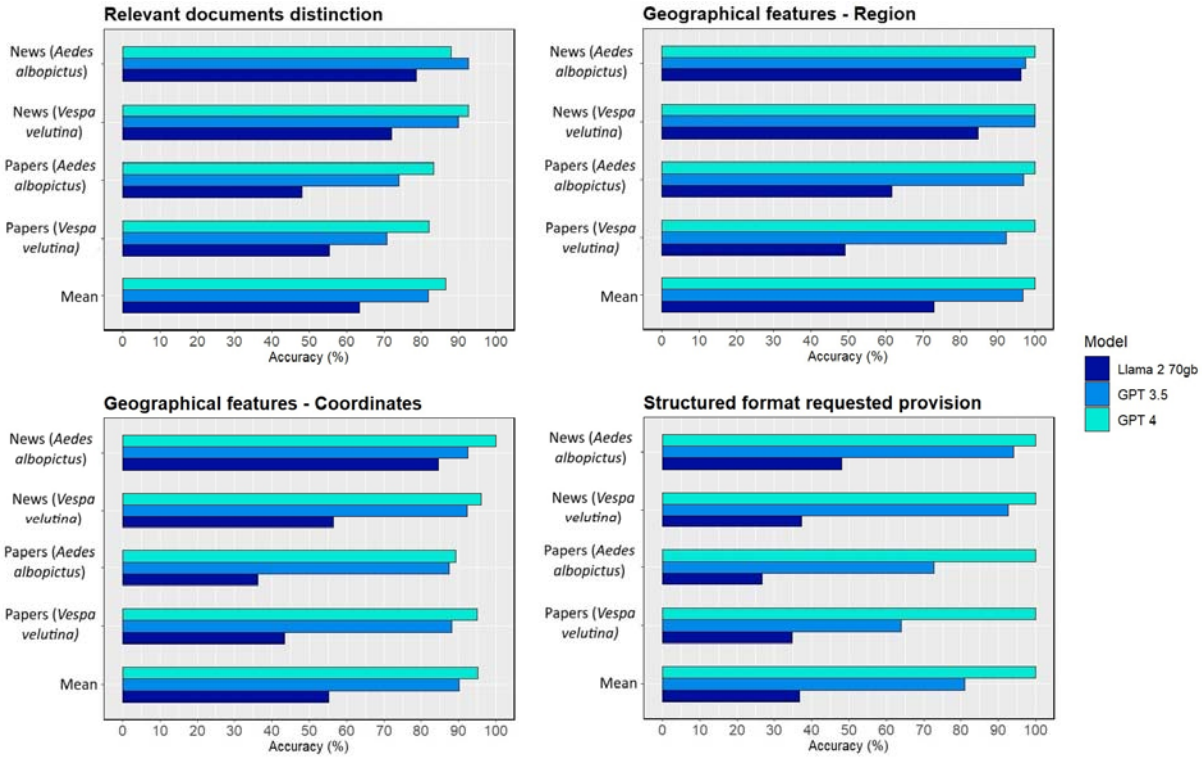

Figure 1 – Accuracy of results for models assessed. These concern: (a) the distinction between documents that are reporting a presence of a given specie and those which do not reporting; (b) the provision of the region(s) name(s) referred in a given document of interest; (c) the provision of coordinates of the region(s) identified; (d) the consistency of models in providing results in a structured format.

Table 3 – Percentage of true positives and true negatives for models classifying between documents reporting the presence of the species and those that do not

		Large language models classification					
		GPT-4		**GPT-3.5**		**Llama 2 70gb**	
Specie name	**Type of Document**	**True positives (%)**	**True negatives (%)**	**True positives (%)**	**True negatives (%)**	**True positives (%)**	**True negatives (%)**
Aedes albopictus	News	75.0	93.4	90.9	93.4	59.1	86.8
	Papers	60.7	98.9	52.5	87.6	77.0	28.1
Vespa velutina	News	96.2	88.7	97.5	81.7	58.2	87.3
	Papers	61.5	97.6	78.5	64.7	81.5	35.3
	Mean	73.35 ± 16.58	94.7 ± 4.61	79.9 ± 19.86	81.9 ± 12.39	69 ± 12.04	54.4 ± 32.09

4 | DISCUSSION

This study underpins the high potential of recent state-of-the-art LLMs in identifying, interpreting, and structuring relevant ecological data from both unstructured and structured textual sources. This marks a significant advancement towards automating the integration of these data into ecological research workflows, potentially alleviating the burden of manual tasks, given that the models can be applied also programmatically via APIs. However, we also observed that the quality of the results varies substantially depending on the type of data source and, particularly, on the LLM used.

Among the three LLMs compared, GPT-4 generally achieved the highest performance across the four evaluated parameters: classification of documents in terms of

distribution data reporting, identification of the regions with distribution data, generation of geographical coordinates, and delivery of results in a structured format. The observed superiority of GPT-4 was expected, being an upgrade of the also-tested GPT-3.5, which has previously shown improvements in several domains, including natural language interpretation (Espejel et al., 2023). It is important to note, however, that for certain types of input and tasks, particularly in producing structured output, GPT-3.5 performs only slightly less effectively. Additionally, GPT models consistently outperformed LLaMA-2-70B in all four tasks, often with substantial gains of 40% or more. The large margin of improvement in our study suggests that LLaMA-2-70B, is not a ready-to-go solution for processing unstructured textual ecological data. This is unfortunate considering its open-source nature, which makes it freely available for academic research. However, due to its open-source nature, LLaMA-2-70B, like many other open-source LLMs (e.g., MPT from MosaicML), can be fine-tuned. In other words, researchers or developers can further train these models for specific tasks (as those we tested), which sometimes significantly improves their performance (Tirumala et al., 2023). Therefore, while our results show that the current pre-trained version of LLaMA-2-70B struggles with ecological textual data, this should not be seen as a definitive incapacity, as fine-tuning could presumably enhance its performance. While such fine-tuning falls beyond the scope of our work, it certainly warrants exploration in future work.

We also noted substantial differences in the models' capacities to handle text from news versus research papers. Across models, text classification, region extraction (the two tasks directly dependent on the provided text) and structured format provision (except GPT-4, where no difference between papers and news was found) consistently showed

better results for news than for papers. This is likely due to the more complex and convoluted narrative often found in research papers, which makes model interpretation more challenging. In contrast, news typically aims to provide easily understandable text with simpler language. This difference in comprehension capacity between research papers and news is less pronounced in GPT-4, especially in geographic region identification, further underscoring the need for large, complex models to understand scientific language. Additional factors may also play a role. Parameters such as the informational content of the examples, or even the order in which they are provided, could have an impact (Zhao et al., 2021). While exploring the precise factors that drive the observed differences is outside the scope of our current work, it represents a relevant avenue for future research. Among other potential outcomes, understanding of the effectiveness of using LLMs for data integration across various types and sources of unstructured text data relevant to ecology would be particularly beneficial.

Despite the aforementioned limitations, the capabilities of GPT models, particularly GPT-4, are noteworthy. To our knowledge, the results these models achieve in identifying, extracting, and structuring ecological information from unstructured textual data surpass any previously available automated solutions. Integrating GPT models into ecological workflows, either programmatically or via no-code approaches, is likely to significantly reduce manual workloads and offers a promising avenue for efficiently harnessing the growing volume of unstructured ecological text data. Additionally, while current open-source Large Language Models (LLMs) appear less accurate, they could become viable alternatives after fine-tuning and potentially overcome some of the limitations imposed by commercial models, such as costs and usage caps. Furthermore, as LLMs continue to evolve, we can expect significant performance improvements in

open-source solutions in the near future. Ultimately, we foresee the imminent use of LLMs in creating a seamless workflow for integrating unstructured text data into ecological analyses.

FUNDING

AC was supported by a grant (PRT/BD/152100/2021) financed by the Portuguese Foundation for Science and Technology (FCT) under MIT Portugal Program. AC and CC acknowledge support from FCT through support to CEG/IGOT Research Unit (UIDB/00295/2020 and UIDP/00295/2020). LR was funded through the FCT contract 'CEECIND/00445/2017' under the 'Stimulus of Scientific Employment—Individual Support' and by FCT 'UNRAVEL' project (PTDC/BIA-ECO/0207/2020). PP acknowledge support from the Czech Science Foundation (project no. 23-07278S).

CONFLICT OF INTEREST STATEMENT

No conflict of interest is declared.

REFERENCES

Aria, M., Cuccurullo, C., Le, T., Choe, J. (2023). Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API. R package version 1.2.3, https://docs.ropensci.org/openalexR/.

Bakran-Lebl, K., Zittra, C., Harl, J., Shahi-Barogh, B., Grätzl, A., Ebmer, D., Schaffner, F., & Fuehrer, H. P. (2021). Arrival of the Asian tiger mosquito, Aedes albopictus (Skuse, 1895) in Vienna, Austria and initial monitoring activities. *Transboundary*

and Emerging Diseases, *68*(6), 3145–3150. https://doi.org/https://doi.org/10.1111/tbed.14169

Barbet-Massin, M., Salles, J. M., & Courchamp, F. (2020). The economic cost of control of the invasive yellow-legged Asian hornet. *NeoBiota*, *55*, 11–25. https://doi.org/https://doi:10.3897/neobiota.55.38550

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., Possingham, H. P., & Lindenmayer, D. B. (2019). Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution*, *6*(239), 1–5. https://doi.org/https://doi.org/10.3389/fevo.2018.00239

Boonman, C. C. F., Serra-Diaz, J. M., Hoeks, S., Guo, W. Y., Enquist, B. J., Maitner, B., Malhi, Y., Merow, C., Buitenwerf, R., & Svenning, J. C. (2024). More than 17,000 tree species are at risk from rapid global change. *Nature Communications*, *15*(166), 1–14. https://doi.org/https://doi.org/10.1038/s41467-023-44321-9

Boulton, D., & Hammersley, M. (2006). Analysis of unstructured data. In R. Sapsford & V. Jupp (Eds.), *Data Collection and Analysis* (pp. 243–259). Sage.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *ArXiv*, *2005.14165*, 1–75. https://doi.org/https://doi.org/10.48550/arXiv.2005.14165

Bullens, L., (2023, July). How the tiger mosquito invaded France and what can be done to stop it. *France 24*. https://www.france24.com/en/environment/20230730-how-the-tiger-mosquito-conquered-france-and-what-can-be-done-to-stop-it

Capinha, C., Marcolin, F., & Reino, L. (2020). Human-induced globalization of insular herpetofaunas. *Global Ecology and Biogeography*, *29*(8), 1328–1349.

https://doi.org/https://doi.org/10.1111/geb.13109

Capinha, C., Seebens, H., Cassey, P., García-Díaz, P., Lenzner, B., Mang, T., Moser, D., Pyšek, P., Rödder, D., Scalera, R., Winter, M., Dullinger, S., & Essl, F. (2017). Diversity, biogeography and the global flows of alien amphibians and reptiles. *Diversity and Distributions*, *23*(11), 1313–1322. https://doi.org/https://doi.org/10.1111/ddi.12617

Carballo, R. (2023, August). An Invasive Hornet Species Is Spotted in the U.S. for the First Time. *The New York Times*. https://www.nytimes.com/2023/08/17/us/yellow-legged-hornet-us.html

Castro, A., Ribeiro, J., Reino, L., & Capinha, C. (2023). Who is reporting non-native species and how? A cross-expert assessment of practices and drivers of non-native biodiversity reporting in species regional listing. *Ecology and Evolution*, *13*(5), 1–11. https://doi.org/https://doi.org/10.1002/ece3.10148

Chiu, K.-L., Collins, A., & Alexander, R. (2021). Detecting Hate Speech with GPT-3. *ArXiv*, *abs/2103.1*, 1–29. http://arxiv.org/abs/2103.12407

Chowdhury, S., Ahmed, S., Alam, S., Callaghan, C., Das, P., Di Marco, M., Di Minin, E., Jarić, I., Labi, M., Rokonuzzamanm, M., Roll, U., Sbragaglia, V., Siddika, A., & Bonn, A. (2023). A standard protocol for harvesting biodiversity data from Facebook. *EcoEvoRxiv*, 1–18. https://doi.org/https://doi.org/10.32942/X2XS4F

Cornford, R., Deinet, S., De Palma, A., Hill, S. L. L., McRae, L., Pettit, B., Marconi, V., Purvis, A., & Freeman, R. (2020). Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. *Global Ecology and Biogeography*, *30*(1), 339–347. https://doi.org/https://doi.org/10.1111/geb.13219

Dillane, E., Hayden, R., O'Hanlon, A., Butler, F., & Harrison, S. (2022). The first recorded occurrence of the Asian hornet (Vespa velutina) in Ireland, genetic

evidence for a continued single invasion across Europe. *Journal of Hymenoptera Research*, *93*, 131–138. https://doi.org/https://doi.org/10.3897/jhr.93.91209

Espejel, J., Ettifouri, E. H., Yahaya Alassan, M. S., Chouham, E. M., & Dahhane, W. (2023). GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, *5*, 1–192. https://doi.org/https://doi.org/10.1016/j.nlp.2023.100032

Frie, P., Yannik, B., Lars, S., Jan, D. (2019). Client for the News API. R package version 0.1.1, https://newsapi.org/.

Grossi, V., Giannotti, F., Pedreschi, D., Manghi, P., Pagano, P., & Assante, M. (2021). Data science: a game changer for science and innovation. *International Journal of Data Science and Analytics*, *11*(4), 263–278. https://doi.org/https://doi.org/10.1007/s41060-020-00240-2

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162. https://doi.org/https://doi.org/10.1890/120103

Hart, A. G., Carpenter, W. S., Hlustik-Smith, E., Reed, M., & Goodenough, A. E. (2018). Testing the potential of Twitter mining methods for data acquisition: Evaluating novel opportunities for ecological research in multiple taxa. *Methods in Ecology and Evolution*, *9*(11), 2194–2205. https://doi.org/https://doi.org/10.1111/2041-210X.13063

Kheiri, K., & Karimi, H. (2023). SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning. *ArXiv*, *abs/2307.1*, 1–14. https://doi.org/https://doi.org/10.48550/arXiv.2307.10234

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, *535*, 457–458. https://doi.org/https://doi.org/10.1038/nj7612-457a

Latombe, G., Pyšek, P., Jeschke, J. M., Blackburn, T. M., Bacher, S., Capinha, C., Costello, M. J., Fernández, M., Gregory, R. D., Hobern, D., Hui, C., Jetz, W., Kumschick, S., McGrannachan, C., Pergl, J., Roy, H. E., Scalera, R., Squires, Z. E., Wilson, J. R. U., … McGeoch, M. A. (2017). A vision for global monitoring of biological invasions. *Biological Conservation*, *213*, 295–308. https://doi.org/https://doi.org/10.1016/j.biocon.2016.06.013

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/https://doi.org/10.1111/2041-210X.13778

Mao, R., Chen, G., Zhang, X., Guerin, F., & Cambria, E. (2023). GPTEval: A Survey on Assessments of ChatGPT and GPT-4. *ArXiv*, *2308.12488*, 1–18. https://doi.org/https://doi.org/10.48550/arXiv.2308.12488

Maquart, P., Fontenille, D., Rahola, N., Yean, S., & Boyer, S. (2021). Checklist of the mosquito fauna (Diptera, Culicidae) of Cambodia. *Parasite*, *28*(60), 1–24. https://doi.org/https://doi.org/10.1051/parasite/2021056 RESEARCH

Moloney, G. K., Tuke, J., Grande, E. D., Nielsen, T., & Chaber, A. L. (2021). Is YouTube promoting the exotic pet trade? Analysis of the global public perception of popular YouTube videos featuring threatened exotic animals. *PLoS ONE*, *16*(4), 1–16. https://doi.org/https://doi.org/10.1371/journal.pone.0235451

Monceau, K., & Thiery, D. (2016). Vespa velutina - current situation and perspectives. *Atti Accademia Nazionale Italiana Di Entomologia*, 137–142.

Monteiro, M., Reino, L., Schertler, A., Essl, F., Figueira, R., Teresa, M., & Capinha, C.

(2020). A database of the global distribution of alien macrofungi. *Biodiversity Data Journal*, *8*(e51459), 1–13. https://doi.org/https://doi.org/10.3897/BDJ.8.e51459

Mota, J. A. R., Silva, J. H. A., Resendes, H. E. M., & Medeiros, N. C. A. (2006). *Popillia Japonica Newman: Relatório dos trabalhos efectuados em 2006 e propostas de actuação para 2007*. https://agricultura.azores.gov.pt/wp-content/uploads/2021/10/relatriopjaponica20092.pdf

OpenAI. (2023). GPT-4 Technical Report. *ArXiv*, *2303.08774*, 1–100. https://doi.org/https://doi.org/10.48550/arXiv.2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Sandhini, A., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *ArXiv*, *abs/2203.0*, 1–68. https://doi.org/https://doi.org/10.48550/arXiv.2203.02155

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., & Peres-Neto, P. (2019). Ecological Data Should Not Be So Hard to Find and Reuse. *Trends in Ecology and Evolution*, *34*(6), 494–496. https://doi.org/10.1016/j.tree.2019.04.005

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *Homology, Homotopy and Applications*, *9*(1), 399–438. https://doi.org/10.4310/HHA.2007.v9.n1.a16

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. 1–24. http://arxiv.org/abs/2007.07582

Schaffner, F., & Mathis, A. (2014). Dengue and dengue vectors in the WHO European region: Past, present, and scenarios for the future. *The Lancet Infectious Diseases*, *14*(12), 1271–1280. https://doi.org/https://doi.org/10.1016/S1473-3099(14)70834-5

Schaffner, F., Medlock, J. M., & Van Bortel, W. (2013). Public health significance of invasive mosquitoes in Europe. *Clinical Microbiology and Infection*, *19*(8), 685–692. https://doi.org/https://doi.org/10.1111/1469-0691.12189

Sethi, S. S., Jones, N. S., Fulcher, B. D., Picinali, L., Clink, D. J., Klinck, H., Orme, C. D. L., Wrege, P. H., & Ewers, R. M. (2020). Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(29), 17049–17055. https://doi.org/https://doi.org/10.1073/pnas.2004702117

Tirumala, K., Simig, D., Aghajanyan, A., & Morcos, A. S. (2023). D4□: Improving LLM Pretraining via Document De-Duplication and Diversification. *ArXiv*, *2308.12284*. https://doi.org/https://doi.org/10.48550/arXiv.2308.12284

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, *abs/2302.1*, 1–27. http://arxiv.org/abs/2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, *aabs/2307*. https://doi.org/https://doi.org/10.48550/arXiv.2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv*, *1706.03762*.

https://doi.org/https://doi.org/10.48550/arXiv.1706.03762

Vuorisalo, T., Lahtinen, R., & Laaksonen, H. (2001). Urban biodiversity in local newspapers: A historical perspective. *Biodiversity and Conservation*, *10*, 1739–1756. https://doi.org/https://doi.org/10.1023/A:1012099420443

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. ArXiv, 2102.09690, 1–15

APPENDIX 1 - Database of news and papers

APPENDIX 2 - Full text prompt

Available from: https://doi.org/10.5281/zenodo.10558495