Check for updates

Received: 23 January 2024 | Accepted: 12 April 2024

DOI: 10.1111/2041-210X.14341

Methods in Ecology and Evolution

RESEARCH ARTICLE

Large language models help facilitate the automated synthesis of information on potential pest controllers

Daan Scheepens[1] | Joseph Millard[2] | Maxwell Farrell[3] | Tim Newbold[1]

[1]Centre for Biodiversity and Environment Research, University College London, London, UK

[2]Biodiversity Futures Lab, Natural History Museum, London, UK

[3]Centre for Virus Research, University of Glasgow, Glasgow, UK

Correspondence
Daan Scheepens
Email: daan.scheepens.22@ucl.ac.uk

Funding information
Natural Environment Research Council, Grant/Award Number: 178836

Handling Editor: Robin Boyd

Abstract

1. The body of ecological literature, which informs much of our knowledge of the global loss of biodiversity, has been experiencing rapid growth in recent decades. The increasing difficulty of synthesising this literature manually has simultaneously resulted in a growing demand for automated text mining methods. Within the domain of deep learning, large language models (LLMs) have been the subject of considerable attention in recent years due to great leaps in progress and a wide range of potential applications; however, quantitative investigation into their potential in ecology has so far been lacking.
2. In this work, we analyse the ability of GPT-4 to extract information about invertebrate pests and pest controllers from abstracts of articles on biological pest control, using a bespoke, zero-shot prompt.
3. Our results show that the performance of GPT-4 is highly competitive with other state-of-the-art tools used for taxonomic named entity recognition and geographic location extraction tasks. On a held-out test set, we show that species and geographic locations are extracted with F1-scores of 99.8% and 95.3%, respectively, and highlight that the model can effectively distinguish between ecological roles of interest such as predators, parasitoids and pests. Moreover, we demonstrate the model's ability to effectively extract and predict taxonomic information across various taxonomic ranks. However, we do report a small number of cases of fabricated information (confabulations).
4. Due to a lack of specialised, pre-trained ecological language models, general-purpose LLMs may provide a promising way forward in ecology. Combined with tailored prompt engineering, such models can be employed for a wide range of text mining tasks in ecology, with the potential to greatly reduce time spent on manual screening and labelling of the literature.

KEYWORDS
biocontrol, ChatGPT, GPT-4, large language model, prompt engineering, relation extraction, taxonomic named entity recognition, text mining

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

1 | INTRODUCTION

Much of our knowledge of the global loss of biodiversity stems from large-scale syntheses of the ecological literature, such as the WWF Living Planet Index (LPI, 2024), PREDICTS (Hudson et al., 2017), and BioTIME (Dornelas et al., 2018) databases, as well as global reports such as those of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES, 2019). The ecological literature has simultaneously seen rapid growth in recent decades (Anderson et al., 2021) making it increasingly difficult to synthesise this literature manually (Ananiadou et al., 2009). Thus, there is a growing demand for automated methods. Text mining and natural language processing (NLP) methods are expected to have significant potential in automating tasks such as document classification, named entity recognition (NER) and disambiguation, and the extraction of relations between entities (Farrell et al., 2022). Previous approaches in ecology have focused heavily on named entity recognition of species and taxonomy (Akella et al., 2012; Gerner et al., 2010; Le Guillarme & Thuiller, 2022; Millard et al., 2020) as well as geographical locations or population trends (Cornford et al., 2022) but also include document classification (Cornford et al., 2021) and relation extraction (Kaur et al., 2019). Furthermore, various gold standard databases of species names and taxonomy have been published to aid the evaluation of NER approaches in ecology (Abdelmageed et al., 2022; Nguyen et al., 2019).

Within the domain of deep learning (DL) for NLP tasks, large language models (LLMs) have been the subject of considerable attention in recent years by virtue of great leaps in progress and a wide range of potential applications (OpenAI, 2023a). This deep learning revolution in NLP is driven by the transformer architecture (Vaswani et al., 2023), which underlies many innovative DL tools in the natural sciences, such as the AlphaFold model for protein structure prediction (Jumper et al., 2021). Recent transformer-based LLMs are trained on large amounts of input data such that they are able to generate realistic text and facilitate human-computer interaction via natural language (Ouyang et al., 2022). Provided with the right prompts, these models can exhibit advanced reasoning and problem-solving capabilities (Kojima et al., 2023; Wei et al., 2023; Zhou et al., 2023). In particular, OpenAI's fourth-generation Generative Pre-trained Transformer (GPT-4) has seen major improvements over previous models on a variety of benchmarks (OpenAI, 2023a). GPT-based models have already been applied to multiple research domains, including finance (Wu et al., 2023) and medical science (Chen et al., 2023; Hu et al., 2023; Silva et al., 2024), however, quantitative investigation into their potential use ecology is lacking.

GPT-4 is a current state-of-the-art large language model and is easily accessed using ChatGPT, making it an attractive choice to demonstrate the potential of LLMs for automated information extraction and knowledge synthesis from scientific texts. However, it is important to note that there exist many competitive open source alternatives to closed-access models such as OpenAI's GPT series, and the landscape of open-source LLMs is rapidly evolving (e.g. Dey et al., 2023; Touvron et al., 2023; Zhang et al., 2022).This availability of open-source LLMs is crucial to fostering open, reproducible and equitable uses of AI in ecology. Thus, while GPT-4 does not currently adhere to open science standards, the approach outlined in this paper is serves as a proof of concept for the use of prompt-based interaction with general-purpose LLMs for ecological information extraction.

Here we use GPT-4 to extract information from ecological literature on biological pest control. Natural enemies of pests, such as arthropod predators and parasitoids, can be used as biological control agents and provide an effective way to reduce pesticide usage, which is currently a major driver of insect declines (Wagner et al., 2021). Biological control has historically relied on the introduction of non-native species (classical biocontrol), which can be detrimental to local ecosystems and negatively affect biological diversity. Natural biological control, on the other hand, utilises native species as biological control agents and is typically achieved through the incorporation (and enhancement) of natural habitat in agricultural systems. Natural biological control helps directly regulate the frequency of pest outbreaks (Letourneau, 2012) and, indirectly—as a result of reduced pesticide usage and increased natural habitat—can result in improved soil quality (Gunstone et al., 2021), increased crop yields (Dainese et al., 2019) and increased abundances of other beneficial organisms such as pollinators (Wratten et al., 2012). Crucially, by strengthening the stability and resilience of ecosystem services such as pest control, pollination and nutrient cycling, food production systems can be better buffered against environmental and climatic changes (Brittain et al., 2013; Martin et al., 2019; Oliver et al., 2015), which is of growing importance as the impacts of climate change continue to intensify (IPCC, 2023).

Here we analyse the ability of GPT-4 to reliably extract information from scientific abstracts to identify pests and pest controllers. In addition to determining the ecological roles of species, the model is tasked with extracting taxonomy, geographic location and role-specific information such as pest type and the crop or plant that a pest affects. As such, the task comprises multiple subtasks, which require both the capability to recognise and disambiguate entities (named entity recognition), and to extract relations between entities (relation extraction). Rather than fine-tuning the parameters of the model, we proceed to optimise model performance through the fine-tuning of the prompt itself. Finally, performance is analysed for each subtask of the query with the help of proportion-correct, precision, recall and F1-scores. To our knowledge, this is the first instance of a general-purpose, large language model such as GPT-4 being used for the automation of information extraction and knowledge synthesis in ecology.

2 | METHODS

2.1 | Data collection

In order to obtain relevant literature on potential pest controllers and their hosts, we extracted a set of abstracts from the academic

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

indexing tool Scopus up until the year 2020, using the following search term:

> TITLE-ABS-KEY("pest control" OR "biological control" OR "pest management" OR "natural enem*") AND (LIMIT-TO(DOCTYPE, "ar")) AND (LIMIT-TO(SUBJAREA, "AGRI") OR LIMIT-TO(SUBJAREA, "ENVI")) AND (LIMIT-TO(LANGUAGE, "English"))

The usage of Scopus ensures that all of the extracted literature has undergone peer-review, which is important, given the underlying motive of this work is to automate knowledge synthesis for peer-reviewed manuscripts. Our search yielded a corpus of 58,791 abstracts, from which we selected a subset of 100 abstracts to create a training set, used to fine-tune the prompt, and a further 100 abstracts to create a held-out test set, used to analyse the final performance of GPT-4 with the fine-tuned prompt. We highlight that the term 'training' thus does not refer to the training of model parameters, as it is more commonly used to refer to in the machine learning literature. We populated both of these sets by manually labelling the 200 subset abstracts (including titles and keywords) using the columns shown in Table 1, including all species found at the species or genus level, but excluding plants, bacteria, fungi and pathogens. Thus, we focus here on pest control services of, and provided by, animal species specifically. The manual labelling was carried out by one annotator (Daan Scheepens), based on the definitions shown in Table 1. Species were labelled based only on information in the abstract, with the only exception being biological control agents that lacked an explicit role but could easily be inferred to be either predators or parasitoids based on their taxonomy (e.g. ground beetles or braconid wasps).

To comprise the training set we selected a stratified sample of 100 abstracts to include both predators and parasitoids, in addition to pests, in order to fine-tune the prompt on these roles. The information on instances of predators, parasitoids and pests was available from a prescreening procedure for a total of 1520 abstracts, which identified and described genera present in these abstracts and could thus indicate which abstracts contained (at least one genera) of predators, parasitoids and pests. The 100 abstracts comprising the training set were obtained from this subset one-by-one, with a continuous attempt to balance the number of predators, parasitoids and pests (Figure S1a). The 100 abstracts comprising the held-out test set were selected randomly from the remaining abstracts in the corpus. Pest species are common in the corpus, which means that there is a relative increase in the proportion of species labelled as pests in the test set compared to the training set (Figure S1b).

2.2 | Prompt and objective

We applied GPT-4 to each individual abstract (title, abstract and author keywords) in the training set, with the instruction to find all taxa mentioned at the species or genus level, and to return a table with the columns shown in Table 1. To achieve this, we instructed GPT-4 with an initial prompt (Figure S7), which was consequently fine-tuned against the training set (see Section D of the Supporting Information).

Analogous to the species filtering during manual labelling, we prompted GPT-4 to skip any mentions of plants, bacteria, fungi or pathogens. Although this does not exclude vertebrates per se, we analyse the performance of the model on invertebrates only, as no vertebrates were extracted during manual labelling. While there are cases of vertebrates and other animals groups mentioned only with their common names (e.g. the sheep in 'sheep scab mite'), we discarded these from the predictions as they are typically not organisms of interest to the study and were indeed not manually extracted. The ability of GPT-4 to neglect certain entities in the text is particularly useful for cases where generated output should be limited to certain species of interest. For large-scale studies, the increased completion time of generating exhaustive tables may be substantial, and the generation of comprehensive output may be infeasible due to limitations on token lengths: At the time of this study, both the input and prompt completion of GPT-4 were limited to 2048 tokens (OpenAI, 2023b), corresponding to approximately 1536 words (OpenAI, 2023c).

Following fine-tuning, the prompt assumed the design as shown in Figure 1. By splitting this complex task into three smaller prompts, we adopted what is known as least-to-most prompting (Zhou et al., 2023). The prompt design follows a 'zero-shot' approach, as it contains only instructions rather than including any exemplar prompt completions (which would have drastically increased the token length of the prompt). We attempted to improve the reasoning ability of the model by including the phrase 'let's think through the following tasks step by step' (info-point 3; Kojima et al., 2023), and through the usage of chain-of-thought reasoning (info-points'C'; Wei et al., 2023). These prompting techniques have all been demonstrated to improve the reasoning abilities of GPT-3.5 (Kojima et al., 2023; Wei et al., 2023; Zhou et al., 2023). The prompt also makes abundant usage of specific examples and counterexamples to aid in the correct identification of particular species roles. We applied this fine-tuned prompt to the training data once more without any further changes to obtain final results on the training set, and then applied the prompt to the held-out test set.

2.3 | Evaluation

As shown in Table 1, the first five columns of the datasets contain the taxonomy of the species or genus. While higher-level taxonomic information can be extracted from databases such as the Global Biodiversity Information Facility (GBIF, 2023) for a given species or genus, species names are frequently misspelled in the literature and genus names occasionally appear across different phyla, making it difficult to automatically extract the correct taxonomy in all instances. We therefore deemed it of interest to investigate the ability

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Columns of the data used in this study.

	Column	Explanation	Clarification
1.	Class	Class name	Taxonomy (Latin)
2.	Order	Order name	Taxonomy (Latin)
3.	Family	Family name	Taxonomy (Latin)
4.	Genus	Genus name	Taxonomy (Latin)
5.	Species	Species name	Taxonomy (Latin)
6.	Role	The role that the species is ascribed in the text (e.g. predator, parasitoid, pest, herbivore, etc.)	Not limited to the given examples
7.	G/S	Is the species mentioned as a generalist or specialist?	This refers to dietary generalism and specialism
8.	Pest Controller	True or false: Is the species mentioned as a pest controller?	We define a pest controller as a natural enemy (i.e. predator or parasitoid) of a pest
9.	Pest Names	For pest controllers: What pests does it control?	May be given as a common name or at any taxonomic scale, as mentioned in the abstract
10.	Pest Type	For pests and pest controllers: Is the pest an invertebrate or plant (i.e. weed)?	Although plant species are not to be returned as species in the table, they may be mentioned in Column 9
11.	Industry Type	For pests and pest controllers: What industry is the pest associated with (e.g. agriculture, forestry, freshwater, etc.)?	Not limited to the given examples
12.	Affects	For pests and pest controllers: What crops, plants or products does the pest affect?	This may be given as common names or at any taxonomic scale, as mentioned in the abstract
13.	Description	Thorough description of the species as mentioned in the abstract and a justification for the assessment	Unconstrained in length
14.	Location	Geographical location of the study, if mentioned	Name of the location as mentioned in the abstract, unconstrained in scale

Note: Using these columns, we manually populated the training set (using the 100 abstracts selected for the training set) and the test set (using the 100 abstracts selected for the test set): These constitute the 'manually labelled data' in the training and test set. Consequently, we instructed GPT-4 to generate tables with the same columns: These constitute the 'GPT-4 generated' training and test set.

of GPT-4 to (1) predict the missing higher-level taxonomy of a given species or genus from its training data and (2) correct any suspected misspellings.

We refer to the task of extracting taxonomy that is stated in the abstract as 'taxonomic named entity recognition'. Where the taxonomy was not available in the abstract, we searched for the species in GBIF and the Encyclopedia of Life (EOL) (Parr et al., 2014) using the genus and (if available) species name and filled this out accordingly. We refer to GPT-4's task of predicting this missing taxonomy as 'higher-level taxonomy prediction'. This taxonomy concerns only the class, order and family columns in the data sets, as the genus and species columns are needed to predict the missing taxonomy. In both cases, the obtained taxonomic information is then compared with the generated GPT-4 output. Since taxonomic named entity recognition and higher-level taxonomy prediction comprise two very different tasks, we evaluate performance of the model on these two groups of taxonomic terms individually. We do note that GPT-4 does not inherently distinguish between these two tasks, but rather generates taxonomic terms simply on the basis of the information available at any instance (which includes both the text at hand and the information learned from its training data). Therefore, given a generated table, it is not possible to say a posteriori which higher-level taxonomic terms have been extracted and which have been predicted from the model's training data.

Based on our definitions of exact and approximate matches (see Section E of the Supporting Information), we evaluate the extraction (and prediction) of taxonomy with the proportion correct (PC) score, which computes the ratio of correct predictions to the total number of predictions. We evaluate this score both against exact matches (i.e. counting all mismatches as mistakes) and against approximate matches (i.e. counting only major mistakes as mistakes):

$$\begin{aligned} PC_{exact} &= \frac{n_{exact}}{N} = \frac{N - \text{mismatches}}{N} \\ PC_{approx} &= \frac{n_{approx}}{N} = \frac{N - \text{majormistakes}}{N} \end{aligned} \tag{1}$$

where N is the total number of predictions, n_{exact} is the number of exact matches and n_{approx} is the number of approximate matches. We focus our analysis of the remaining species information (columns 6–14) on the species roles and the geographic locations. For

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Methods in Ecology and Evolution

FIGURE 1 The final prompt design after fine-tuning against the training set. The prompt is divided into three parts. The first part prompts GPT-4 to detect all relevant species in the abstract, the second part prompts GPT-4 to return these in the table with the appropriate columns filled out, and the third part prompts GPT-4 to review the generated table and make any corrections if necessary. These three prompts were always executed within the same session. Points labelled as'C' (and highlighted in grey rather than yellow) designate clarifications; these were used to clarify specific terms such as pest and pest controller, and aimed to reduce persistent problems such as formatting errors and missing pest information in some columns.

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

these two subtasks, we measure the performance of GPT-4 by computing the precision, recall and F1-score (Equation 2) of the model based only on approximate matches (Section E of the Supporting Information). The F1-score is computed as the harmonic mean (i.e. the reciprocal of the arithmetic mean of the reciprocals) of precision and recall, making it a function of both measures. Unlike the arithmetic mean, the harmonic mean is 0 when either recall or precision are 0. Although this is also true for a geometric mean, a harmonic mean further penalises uneven performance.

$$\begin{gathered}\text{Precision} = \frac{TP}{TP+FP} \\ \text{Recall} = \frac{TP}{TP+FN} \\ \text{F1-¬score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \\ TP\text{: True Positives} \\ FP\text{: False Positives} \\ FN\text{: False Negatives}\end{gathered} \tag{2}$$

3 | RESULTS

3.1 | Species extraction

We found that splitting the instructions into two consecutive prompts, in which the first prompt served to identify all relevant species in the text and the second prompt served to fill out the table, led to improvement in entity extraction capability. Although the initial prompt already extracted 627 out of 649 species (96.6% recall) in the training set, the split-prompt approach ensured that all 649 species were extracted in the first part of the prompt. However, in the second part of the prompt, GPT-4 remained challenged to return all previously found species in the final table, and as a result, a small number of species in the training data were still missed in GPT-4's output, with 631 out of 649 species extracted (97.2% recall) (Table 2a). However, since these missing entries correspond to only 2 out of 100 abstracts in the training set, this obtains a mean recall *per abstract* of 99.5% (standard deviation = 4.4%; Table 2b).

Importantly, we also observed 43 instances of fabricated species entries (i.e. 'confabulations') as rows in the GPT-4 output for the training set, which occurred in the same two abstracts where missing entries were observed. These false positives result in a precision of 93.6% and a mean precision per abstract of 99.1% (standard deviation = 6.8%; Table 2) for the training set. For the test set, GPT-4 extracted 244 of 245 species (99.6% recall), and we did not observe any confabulated species entries (100% precision) (Table 2). Moreover, we found that GPT-4 did not generate any entries for plants, bacteria, fungi or pathogens in either the training set or the test set and thus managed to abide by the constraint set on the species extraction very effectively.

We hypothesised that the third step in the final prompt design, in which we instructed GPT-4 to thoroughly review its output, would allow the model to pick up on its own mistakes; however, its success was found to be very limited. GPT-4 corrected its own mistakes only in the case of a single abstract in the training set, corresponding to a table with 12 species rows. In this case, the model correctly identified that it had left out columns 10, 11 and 12 for the pest species in the table and then proceeded to fill these out correctly. In one case in the test set, the model claimed that it had not identified a particular species as 'pest' and stated that it had corrected this, although the species had, in fact, already been identified as a pest. In two other cases, the model returned reassurances other than 'No corrections' (although conveying the same message). The model did not recognise missed species or confabulations.

The generated tables obtained from GPT-4 display varying degrees of accuracy and adherence to the prompted instructions. However, in cases where the information in the abstract is presented clearly and nonambiguously, strong performance can be observed (Figure S2). In contrast, ambiguous usage of language (e.g. a predatory species that acts as a pest) is prone to produce erroneous results (Figure S3). In the remaining analysis, the species in the training and test set that were missed by GPT-4 are omitted as these species naturally cannot be compared to the manual labels. This refers to 18 out of 642 species in the training set and 1 out of 245 species in the test set. For the same reason, we also omit the 43 cases of confabulations in the training set from this further analysis, but discuss these at length in the discussion.

3.2 | Taxonomic named entity recognition

We observed only a small number of mismatches between manual labels and GPT-4 predictions in the case of taxonomic named entity recognition, resulting in high PC_{exact} scores (Table 3a). Following comparison with the reference taxonomy, we found that the majority of mismatches

	Precision (%)	Recall (%)	F1 (%)	Support
(a) Total				
Training set	93.6	97.2	95.3	649
Test set	100.0	99.6	99.8	245
(b) Mean per abstract				
Training set	99.1 ± 6.8	99.5 ± 4.4	99.3 ± 4.1	100
Test set	100.0 ± 0.0	99.0 ± 9.9	99.5 ± 5.0	100

Note: Support: The number of species present in the manually labelled dataset (a); the number of abstracts in the dataset (b).

TABLE 2 Precision, recall and F1-score of the extraction of species (as rows in the final output table) by GPT-4, for the training set and the held-out test set. Scores are presented as computed from the total number of predicted and manually extracted species entries (a) and as the mean per abstract (b).

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

were either cases where the GPT-4 prediction was, in fact, correct (23/44 in the training set and 23/26 in the test set), or where the mistake comprised only a minor, rather than a major mistake (Table S2a). Indeed, we observed only one major mistake in the 2240 total taxonomic terms (0.04%) predicted by GPT-4 in the training set and three major mistakes in the 801 total taxonomic terms (0.37%) predicted by GPT-4 in the test set (Table S2a). We also note that PC_{approx} scores amount to 100% for the genus rank, and are in excess of 99.6% for the species rank, in both sets (Table S3a). These ranks bear particular importance, as any higher-level taxonomy obtained from databases such as GBIF depends on their correctness. GPT-4 does obtain similar performance for the class, order and family ranks, however, suggesting that when higher-level taxonomy is mentioned in the text, it is extracted reliably.

We note that mismatches where the GPT-4 prediction turned out to be correct refer primarily to cases of spelling corrections in the genus and species columns, and cases of corrected ranks (e.g. the correct family as opposed to the superfamily mentioned in the abstract) in the class, order and family columns, although we also find spelling corrections in the family column. A common case in the order column comprises the manually extracted taxa 'Homoptera' and 'Heteroptera', both of which are suborders of Hemiptera, although the former is now considered obsolete by many taxonomists (von Dohlen & Moran, 1995). In terms of spelling corrections, we observed 12 cases in the training set and 4 in the test set. In all cases, GPT-4 corrected the spelling mistake and returned the correct term. While this does not evidence that all misspellings in the data were corrected by the model, it does demonstrate that when a misspelling was addressed, it was done correctly.

3.3 | Higher-level taxonomy prediction

For the task of higher-level taxonomy prediction, we observed more mismatches than for the task of taxonomic NER (as proportions of the total predictions), resulting in generally lower PC_{exact} scores (Table 3b). Furthermore, less of these mismatches comprised cases where the GPT-4 prediction turned out to be correct, and more of the mismatches comprised major mistakes (Table S2b). We observed 15 major mistakes out of the 915 total taxonomic terms (1.64%) predicted by GPT-4 in the training set and 7 major mistakes out of the 419 total taxonomic terms (1.67%) predicted by GPT-4 in the test set (Table S2a). We find that while PC_{approx} scores for the class and order ranks fall between 98%–100% (Table S3b), we observe a notable decline for the family rank in both the training set and the test set. This decline, however, appears to be a result of bias toward a small number of abstracts in which many major mistakes were made and is thus not observed in the scores per abstract (PC^*_{approx}).

3.4 | Species roles

GPT-4 captured the roles of species with a high degree of precision and recall for most labels (Table 4). Crucially, while there are a small number of cases of manually labelled predators, parasitoids and pests that have been mislabelled by GPT-4 as something else (e.g. as 'competitor', 'prey', 'host' or 'unclear'), there is no confusion between these roles, either in the training set or in the test set (Figure S4).

Results for the test set are very comparable to results for the training set, although some differences in precision and recall can be observed (Table 4). In particular, the precision for predator roles is reduced from 100% on the training set to 80.4% on the test set. This reduction is a result of GPT-4 predictions of 'predator' for terms that were manually labelled as biological control agents, natural enemies or were left blank (Figure S4b). Recall of predator roles is only marginally reduced on the test set compared to the training set. In both cases, false negatives are the result of a relatively small percentage of abstracts (Table S4).

Performance for parasitoid roles on the test set closely matches performance on the training set, with precision and recall scores between 97.9%–100% and F1-scores of 98.7%–99.0% (Table 4). In both sets, hyperparasitoids (parasitoids of other parasitoids) are accurately distinguished. For pest roles, we observe a substantial amount of confusion between pests and 'prey/host', 'other' and 'unclear', in both the training and the test set (Figure S4). As a result, we observe a relatively large variation in precision and recall for this role between the training set and the test set (Table 4).

3.5 | Geographic locations

GPT-4 is capable of effectively extracting geographic information, with locations predicted with 98.7% precision and 97.1% recall on the training set and 95.3% precision and recall on the test set (Table 5). Per abstract, an average of 96.3% of manually labelled

TABLE 3 Proportion correct (PC) scores of GPT-4 on (a) taxonomic named entity recognition, (b) higher-level taxonomy prediction and (c) all taxonomic terms, totalled over all ranks. Individual scores per rank can be found in Table S3.

	PC_{exact} (%)	PC_{approx} (%)	PC^*_{approx} (%)
(a) Taxonomic named entity recognition			
Training set	98.0	99.9	99.9 ± 1.2
Test set	96.8	99.6	98.8 ± 10.1
(b) Higher-level taxonomy prediction			
Training set	97.6	98.4	98.1 ± 8.0
Test set	94.5	98.3	96.9 ± 15.3
(c) All taxonomic terms			
Training set	97.9	99.5	99.0 ± 5.8
Test set	96.0	99.2	97.9 ± 13.0

Note: PC scores are presented as measured against exact matches (PC_{exact}) and as measured against approximate matches (PC_{approx}). We also present the latter *per-abstract* (PC^*_{approx}) to avoid bias from abstracts with a large number of extracted species (presented as the mean across abstracts with a spread of one standard deviation).

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Precision and recall obtained by GPT-4 for each role in the training set (a) and the held-out test set (b).

	Precision (%)	Recall (%)	F1 (%)	Support
(a) Training set				
Predator	**100.0**	**93.2**	**96.5**	207
Parasitoid	**99.5**	**97.9**	**98.7**	195
Pest	**95.3**	**81.2**	**87.7**	149
B.C.A.	88.9	100.0	94.1	16
H. Parasitoid	100.0	92.9	96.3	14
Nat. enemy	NA	NA	NA	0
Prey/host	14.3	50.0	22.2	4
Other	87.5	91.3	89.4	23
Unclear	24.0	52.2	32.9	23
Weighted average	94.7	90.2	92.0	631
(b) Test set				
Predator	**80.4**	**89.1**	**84.5**	46
Parasitoid	**100.0**	**98.1**	**99.0**	52
Pest	**87.4**	**96.0**	**91.5**	101
B.C.A.	87.5	63.6	73.7	11
H. Parasitoid	100.0	100.0	100.0	2
Nat. enemy	NA	0.000	NA	3
Prey/host	60.0	42.9	50.0	7
Other	27.3	42.9	33.3	7
Unclear	40.0	13.3	20.0	15
Weighted average	82.4	84.4	82.7	244

Note: Support: The number of instances of the respective role in the manual labels. Weighted average: The average of the respective column, weighted by the support of each row. 'B.C.A.': Biological Control Agent. 'H. Parasitoid': Hyperparasitoid. 'Nat. Enemy': Natural enemy. 'Other' includes terms such as 'Pollinator', 'Herbivore', 'Leaf miner', 'Scavenger', 'Ectoparasite' and 'Competitor' and the term 'Unclear' includes 'Not mentioned', as well as empty entries.
The primary roles of interest (predator, parasitoid and pest) are emphasised in boldface.

TABLE 5 Precision and recall obtained by GPT-4 for geographic locations in the training set (a) and the test set (b).

	Precision (%)	Recall (%)	F1 (%)	Support
(a) Training set				
Location	98.7	97.1	97.9	487
No location	90.8	95.8	93.2	144
Weighted average	96.9	96.8	96.9	631
(b) Test set				
Location	95.3	95.3	95.3	149
No location	92.6	92.6	92.6	95
Weighted average	94.3	94.3	94.3	244

Note: Support: The number of locations and non-locations present in the manual labels. Weighted average: The average of the respective column, weighted by the support of each row. 'Location' refers to the location associated with the study of the species and'No location' refers to the case where no location is mentioned in the abstract; if GPT-4 predicted a location that did not correspond to the manually labelled location, this is designated as a false-positive, and if GPT-4 did not predict a location for a manually labelled location, this is designated as a false negative.

locations are correctly returned by the GPT-4 predictions in both sets (Table S5). No-locations (i.e. the abstract mentioned no location for the respective species) are generally predicted only marginally worse, with the exception of a notably lower precision on the training set (Table 5a), which stems from a total of 14 false negative predictions (Figure S5). These false negatives occur in four abstracts, with most (9 out of 14) comprising empty predictions, and the remaining comprising predictions of insufficient

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

information originating from a single abstract (predictions of 'North Island' rather than 'Australia').

4 | DISCUSSION

The results presented here show that GPT-4 possesses an effective ability to (1) extract taxonomic and geographic entities from abstracts, (2) identify the roles of species from their descriptions in the abstract, and (3) extract relations between entities in the abstract, such as between pest controllers and pests, and pests and certain industries as well as their affected products. However, a number of observations and caveats merit discussion. In addition, a discussion on the problem of ambiguity and its impact on mismatches between species roles is provided in Section F, and a user guide for researchers interested in working with LLMs is provided in Section G of the Supporting Information.

4.1 | Confabulations

The vast pool of information that GPT-4 is able to synthesise appears to provide a notable strength, enabling the model to recognise common usages of language in ecology, such as the conventions regarding taxonomic ranks and the meaning of domain-specific vocabulary. Moreover, it allows the model to make predictions that go beyond merely the information provided in the text, allowing it to correct spelling mistakes, disentangle ambiguities and provide appropriate roles of species based on descriptions in the text.

However, the ability of GPT-4 to synthesise vast quantities of data can also be a source of confabulations. This refers to cases of believable but fabricated information that chat-based large language models such as ChatGPT have been observed to generate (Azamfirei et al., 2023). We observed a total of 43 confabulated species as rows in the output of GPT-4 (all of these in the training set). While the first part of the prompt (Figure 1) consistently returned the correct set of species, a subsequent mismatch was observed between this set and the species returned in the final table. This was observed exclusively in cases where GPT-4 was unable to complete the entire table in one prompt completion and thus required multiple completions to finish the table (using the 'continue generating' function available in the web interface). We suspect that this repeated generation of output may be the cause of these confabulations, possibly as a result of a diminishment of the model's internal memory (Gong et al., 2023). Indeed, the only cases of confabulated species as rows in GPT-4's output were detected in the two longest tables generated by GPT-4 (corresponding to two abstracts in the training set), consisting of 25 and 58 rows, respectively. Furthermore, all 18 species rows that were *missed* by GPT-4 in the training set also originate from these same two abstracts, and thus the same mechanism may be responsible for these false negatives. While confabulations are concerning, it is expected that usage of GPT-4 with higher token-lengths for prompt completion would alleviate this issue. For example, updated versions of GPT-4 are stated to have maximum token-lengths of 8192 and 32,768 tokens respectively, while web-interface usage was limited to 2048 tokens per generation at the time of the experiment (OpenAI, 2023b). The fact that the GPT-4 output for the test set contained no tables larger than 12 rows in length may thus explain why we did not observe instances of confabulated species in the test set.

Additionally, a number of apparent confabulations were observed for the 'Industry Type' column in the generated data sets, where associations of pests with agriculture or forestry were stated without any clear basis in the text. This was observed for 33 out of 631 species in the training set and for 20 out of 244 species in the test set. Upon closer inspection, however, we found that these associations could be corroborated by other Internet sources in the vast majority of cases (Tables S7 and S8). Just seven cases in the training set, corresponding to two abstracts, could not be corroborated by other internet sources and thus appear to constitute actual confabulations.

The apparent inability of GPT-4 to state whether information was drawn from within the text or from outside the text, and if so from where outside the text, is a well-known caveat of the model. However, new plug-ins are continuously being released to address this issue. Connecting GPT-4 to live internet sources, as exemplified by the GPT-4-powered Microsoft Bing Chat, may also provide an effective way to reduce confabulations through the reliable citing of sources. The proneness of large language models to fabricate erroneous, but credible-sounding pieces of information is the subject of increasing discussions in the broader scientific community, with concerns over the accuracy, reliability and accountability of output obtained from LLMs (Birhane et al., 2023). Although AI models are increasingly being used in science and have begun delivering numerous scientific advances (Wang et al., 2023), it is important for scientists to be aware of the limitations of AI tools, such as LLMs and other black-box, deep learning based models and the potential impact of these models on reliability and reproducibility in science.

4.2 | The challenge of prompt design

The question of prompt design (or prompt engineering) poses a challenge to the adoption of large language models in scientific research. Here, we proceeded to optimise the performance of the prompt design against a training set. Our optimisation approach had two main caveats. The first relates to the reliance on manual, trial-and-error-based optimisation, which is time-consuming and may be intractable for very large datasets. The second relates to the problem of overfitting, which, although not well defined in this context (as there is no proper loss function), may arise if the prompt is too tailored to the training set and fails to generalise to the rest of the data. Regardless of the optimisation strategy, however, designing the prompt to take into account the various intricacies of the task (definitions, examples, counterexamples, clarifications) is highly nontrivial and highly

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

data-dependent, as exemplified by the extensive prompt utilised in this work. Furthermore, performance of GPT-4 was observed to be a highly nonlinear function of the prompt design, with small changes in the prompt leading to large changes in the output (for better or worse).

As opposed to extensive prompt engineering, a powerful alternative may be posed by fine-tuning the parameters of the model. Also known as transfer learning (Pan & Yang, 2010), fine-tuning allows users to customise the pretrained machine learning model to their own use cases by retraining the model (or a subset of parameters of the model) on a much smaller, bespoke data set. With many applications in image and text classification (Weiss et al., 2016), transfer learning is already an active field in biomedicine and, to a lesser degree, ecology. OpenAI has released a fine-tuning option for GPT-3.5, which is stated to "match, or even outperform, base GPT-4-level capabilities on certain narrow tasks" (Peng et al., 2023). It is moreover stated to allow for a shortening in prompts by as much as 90% as instructions can be fine-tuned into the model itself. Fine-tuning may thus provide a promising alternative, or addition, to prompt engineering for large language models.

4.3 | Comparison with other tools

To put some of the results obtained in this work into a broader perspective, we provide a comparison with previous studies that have utilised text mining tools for applications in ecology. Previous attempts to extract taxonomic terms from abstracts have resulted in mean recall scores per abstract of 79.5% (Millard et al., 2020) and 93.6% (Cornford et al., 2022) with the help of the R package Taxize (Chamberlain & Szöcs, 2013), while the extraction of geographic locations has been achieved with a mean recall of 82.1% per abstract (Cornford et al., 2022) with the help of the CLIFF-CLAVIN geoparser model (D'Ignazio et al., 2014). Furthermore, an extensive comparison of eight taxonomic NER models over four gold standard ecology corpora (Le Guillarme & Thuiller, 2022) reports scores for approximate matches ranging between 78%–96% (precision), 74%–93% (recall) and 76%–91% (F1-score).

If we can correctly assume the risk of confabulated species extractions to be minimised with sufficiently large token lengths, we may assume such confabulated entries to pose a relatively low risk for future endeavours, which are likely to incorporate longer token lengths. As such, our results on the test set, which did not suffer from confabulated species extractions, may offer a valuable demonstration of the model's potential performance. The extraction of species from abstracts in the test set was achieved with a total precision of 100%, recall of 99.6% and F1-score of 99.8% (Table 2a). Investigating only the taxonomy of the successfully extracted species, we report PC_{approx} scores on the test set of 99.6% for taxonomy that was stated in the abstract (Table 3a), 98.3% for taxonomy that was not stated in the abstract (Table 3b) and 99.2% overall (Table 3c). Geographic locations (as approximate matches) were extracted from the test set with a precision, recall and F1-score of 95.3% (Table 5b), which is only marginally worse than the performance obtained from the training set. We highlight that GPT-4 achieved these results without prior training on these tasks, while simultaneously generating responses to multiple other tasks laid out in the prompt, such as extracting species roles, identifying pest controllers, pest names and associations, and providing detailed descriptions.

5 | CONCLUSIONS

In this work, we explored the potential of the next generation of large language models for the automation of knowledge synthesis. To this end, we investigated a set of abstracts on biological pest control and prompted GPT-4 to extract species, taxonomy and geographical locations, recognise roles, pest-controlling behaviour, pest types, pest associations and mutual relations.

Our results show that GPT-4 is highly capable of this task, with performance on all investigated subtasks largely congruent with the manual labels. Some of the discrepancy appears to be a result of a certain degree of ambiguity in the abstracts and in the task itself. We thus restrained from speaking of ground-truth in this work, since even manual labels are subject to ambiguity and human error: Indeed, we observed a several cases where the predictions of GPT-4 were possibly more accurate than the manual labels. That being said, we also observed confabulations, that is, fabricated text. In the case of confabulated species as rows in the tables, this appears to be a symptom of the token-length limits placed on GPT-4 at the time of the experiment. In the case of individual pieces of fabricated information, it is more difficult to identify root causes, as they may be more dependent on the model parameters and the original training data, and may, in fact, be corroborated by other sources upon closer inspection. We hope that this work makes a valuable contribution to the rapidly evolving domain of the automation of knowledge syntheses in ecology by demonstrating the potential of GPT-4 for this task. Indeed, general-purpose LLMs may provide an interesting way forward in ecology, since there is currently a lack of specialised, pretrained language models. Combined with tailored prompt engineering, LLMs could be used for a broad range of tasks and have the potential to save a large amount of time spent on manual labelling. Through their vast information base, these models can, in principle, be applied to literature spanning many different languages, helping to mitigate the widespread bias of the English language in knowledge syntheses (Amano et al., 2021; Konno et al., 2023)—although it should be noted that current LLM performance on underrepresented languages has been found to be poor (Laskar et al., 2023). Reliability may be enhanced through the integration of live Internet access and a growing number of plug-ins that address the traceability of information. Additionally, an ability to handle full-text papers is likely to be tractable in the near future and further improve the reliability of extracted information. Moreover, using full texts would

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

eliminate the influence of specific journal policies with regards to naming species in the abstract, which may limit LLM performance when using abstracts only.

Finally, we reiterate that although we investigated the application of GPT-4 to literature on biological pest control, the conclusions of this work should be understood in the larger context of the potential of general-purpose LLMs for the automation of knowledge syntheses in ecology. It is important to note, however, that our work does not imply that LLMs are capable of automating the entire evidence synthesis process: A person is still required to delegate tasks, oversee fine-tuning or optimisation, and validate performance. What we have shown here is not yet a set of guidelines, but rather that current state-of-the-art LLMs excel at a number of NLP tasks relevant to large-scale ecological evidence syntheses and meta-analyses. In the spirit of open science, LLM applications ideally transpire in an open-source context, ensuring both open access and reproducibility. Although GPT-4 does not currently adhere to such standards, our aim in this study has been to provide insight into the current state of the art of LLMs and their potential use for automated knowledge synthesis in ecology.

AUTHOR CONTRIBUTIONS

Daan Scheepens and Tim Newbold conceived the ideas and designed methodology; Daan Scheepens collected the data and carried out the manual labelling, analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14341.

DATA AVAILABILITY STATEMENT

All code and data needed to generate the figures in this paper are archived on Zenodo at https://doi.org/10.5281/zenodo.10969983 (Scheepens, 2024).

ORCID

Daan Scheepens https://orcid.org/0009-0007-7574-9621
Joseph Millard https://orcid.org/0000-0002-3025-3565
Maxwell Farrell https://orcid.org/0000-0003-0452-6993
Tim Newbold https://orcid.org/0000-0001-7361-0051

REFERENCES

Abdelmageed, N., Löffler, F., Feddoul, L., Algergawy, A., Samuel, S., Gaikwad, J., Kazem, A., & König-Ries, B. (2022). BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal*, *10*, e89481. https://doi.org/10.3897/BDJ.10.e89481

Akella, L. M., Norton, C. N., & Miller, H. (2012). NetiNeti: Discovery of scientific names from text using machine learning methods. *BMC Bioinformatics*, *13*(1), 211. https://doi.org/10.1186/1471-2105-13-211

Amano, T., Berdejo-Espinola, V., Christie, A. P., Willott, K., Akasaka, M., Báldi, A., Berthinussen, A., Bertolino, S., Bladon, A. J., Chen, M., Choi, C. Y., Bou Dagher Kharrat, M., de Oliveira, L. G., Farhat, P., Golivets, M., Hidalgo Aranzamendi, N., Jantke, K., Kajzer-Bonk, J., Kemahlı Aytekin, M. Ç., … Sutherland, W. J. (2021). Tapping into non-English-language science for the conservation of global biodiversity. *PLoS Biology*, *19*(10), e3001296.

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review*, *27*, 509–523. https://doi.org/10.1177/0894439309332293

Anderson, S. C., Elsen, P. R., Hughes, B. B., Tonietto, R. K., Bletz, M. C., Gill, D. A., Holgerson, M. A., Kuebbing, S. E., McDonough MacKenzie, C., Meek, M. H., & Veríssimo, D. (2021). Trends in ecology and conservation over eight decades. *Frontiers in Ecology and the Environment*, *19*(5), 274–282. https://doi.org/10.1002/fee.2320

Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, *27*(1), 120. https://doi.org/10.1186/s13054-023-04393-x

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, *5*(5), 277–280. https://doi.org/10.1038/s42254-023-00581-4

Brittain, C., Kremen, C., & Klein, A.-M. (2013). Biodiversity buffers pollination from changes in environmental conditions. *Global Change Biology*, *19*(2), 540–547. https://doi.org/10.1111/gcb.12043

Chamberlain, S., & Szöcs, E. (2013). Taxize: Taxonomic search and retrieval in R. *F1000Research*, *2*, 191. https://doi.org/10.12688/f1000research.2-191.v1

Chen, Q., Sun, H., Liu, H., Jiang, Y., Ran, T., Jin, X., Xiao, X., Lin, Z., Niu, Z., & Chen, H. (2023). A comprehensive benchmark study on biomedical text generation and mining with ChatGPT. *bioRxiv* https://doi.org/10.1101/2023.04.19.537463

Cornford, R., Deinet, S., De Palma, A., Hill, S. L. L., McRae, L., Pettit, B., Marconi, V., Purvis, A., & Freeman, R. (2021). Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. *Global Ecology and Biogeography*, *30*(1), 339–347. https://doi.org/10.1111/geb.13219

Cornford, R., Millard, J., González-Suárez, M., Freeman, R., & Johnson, T. F. (2022). Automated synthesis of biodiversity knowledge requires better tools and standardised research output. *Ecography*, *2022*(3), e06068. https://doi.org/10.1111/ecog.06068

Dainese, M., Martin, E. A., Aizen, M. A., Albrecht, M., Bartomeus, I., Bommarco, R., Carvalheiro, L. G., Chaplin-Kramer, R., Gagic, V., Garibaldi, L. A., Ghazoul, J., Grab, H., Jonsson, M., Karp, D. S., Kennedy, C. M., Kleijn, D., Kremen, C., Landis, D. A., Letourneau, D. K., … Steffan-Dewenter, I. (2019). A global synthesis reveals biodiversity-mediated benefits for crop production. *Science Advances*, *5*(10), eaax0121. https://doi.org/10.1126/sciadv.aax0121

Dey, N., Gosal, G., Zhiming, C., Khachane, H., Marshall, W., Pathria, R., Tom, M., & Hestness, J. (2023). Cerebras-GPT: Open compute-optimal language models trained on the cerebras wafer-scale cluster.

D'Ignazio, C., Bhargava, R., & Zuckerman, E. (2014). CLIFF-CLAVIN: Determining geographic focus for news articles [extended abstract].

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., Akhmetzhanova, A. A., Appeltans, W., Arcos, J. M., Arnold, H., Ayyappan, N., Badihi, G., Baird, A. H., Barbosa, M., Barreto, T. E., Bässler, C., Bellgrove, A., Belmaker, J., Benedetti-Cecchi, L., … Hickler, T. (2018). BioTIME: A database of biodiversity time series

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

for the Anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. https://doi.org/10.1111/geb.12729

Farrell, M. J., Brierley, L., Willoughby, A., Yates, A., & Mideo, N. (2022). Past and future uses of text mining in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1975), 20212721. https://doi.org/10.1098/rspb.2021.2721

GBIF. (2023). *GBIF home page*. https://www.gbif.org (visited on 08/25/2023)

Gerner, M., Nenadic, G., & Bergman, C. M. (2010). LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*, *11*(1), 85. https://doi.org/10.1186/1471-2105-11-85

Gong, D., Wan, X., & Wang, D. (2023). *Working memory capacity of ChatGPT: An empirical study*.

Gunstone, T., Cornelisse, T., Klein, K., Dubey, A., & Donley, N. (2021). Pesticides and soil invertebrates: A Hazard assessment. *Frontiers in Environmental Science*, *9*, 1–21. https://doi.org/10.3389/fenvs.2021.643847

Hu, Y., Ameer, I., Zuo, X., Peng, X., Zhou, Y., Li, Z., Li, Y., Li, J., Jiang, X., & Xu, H. (2023). *Zero-shot clinical entity recognition using ChatGPT*.

Hudson, L. N., Newbold, T., Contu, S., Hill, S. L., Lysenko, I., de Palma, A., Phillips, H. R., Alhusseini, T. I., Bedford, F. E., Bennett, D. J., Booth, H., Burton, V. J., Chng, C. W., Choimes, A., Correia, D. L., Day, J., Echeverría-Londoño, S., Emerson, S. R., Gao, D., … Purvis, A. (2017). The database of the PREDICTS (projecting responses of ecological diversity in changing terrestrial systems) project. *Ecology and Evolution*, *7*(1), 145–188. https://doi.org/10.1002/ece3.2579

IPBES. (2019). Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Version 1 https://doi.org/10.5281/zenodo.6417333

IPCC. (2023). Team, C. W., Lee, H., & Romero, J. (Eds.). Sections. In *Climate change 2023: Synthesis report. Contribution of working groups I, II and III to the Sixth Assessment report of the Intergovernmental Panel on Climate Change* (pp. 35–115). IPCC. https://doi.org/10.59327/IPCC/AR6-9789291691647

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaur, K. M., Malé, P.-J. G., Spence, E., Gomez, C., & Frederickson, M. E. (2019). Using text-mined trait data to test for cooperate-and-radiate co-evolution between ants and plants. *PLoS Computational Biology*, *15*(10), 1–18. https://doi.org/10.1371/journal.pcbi.1007323

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large language models are zero-shot reasoners*.

Konno, K., Akasaka, M., Koshida, C., Katayama, N., Osada, N., Spake, R., & Amano, T. (2023). Ignoring non-English-language studies may bias ecological meta-analyses. *Ecology and Evolution*, *10*(13), 6373–6384. https://doi.org/10.1002/ece3.6368

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). *A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets*.

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

Letourneau, D. K. (2012). Integrated pest management—Outbreaks prevented, delayed, or facilitated? In *Insect outbreaks revisited* (pp. 371–394). John Wiley & Sons, Ltd. Chap. 18. https://doi.org/10.1002/9781118295205.ch18

LPI. (2024). *Living planet index*. www.livingplanetindex.org/ (visited on 01/03/2024)

Martin, E. A., Feit, B., Requier, F., Friberg, H., & Jonsson, M. (2019). Chapter three—Assessing the resilience of biodiversity-driven functions in agroecosystems under environmental change. In D. A. Bohan & A. J. Dumbrell (Eds.), *Resilience in complex socio-ecological systems* (Vol. *60*. Advances in Ecological Research, pp. 59–123). Academic Press. https://doi.org/10.1016/bs.aecr.2019.02.003

Millard, J. W., Freeman, R., & Newbold, T. (2020). Text-analysis reveals taxonomic and geographic disparities in animal pollination literature. *Ecography*, *43*(1), 44–59. https://doi.org/10.1111/ecog.04532

Nguyen, N. T., Gabud, R. S., & Ananiadou, S. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, *7*, e29626. https://doi.org/10.3897/BDJ.7.e29626

Oliver, T. H., Isaac, N. J. B., August, T. A., Woodcock, B. A., Roy, D. B., & Bullock, J. M. (2015). Declining resilience of ecosystem functions under biodiversity loss. *Nature Communications*, *6*(1), 10122. https://doi.org/10.1038/ncomms10122

OpenAI. (2023a). *GPT-4 technical report*. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2023b). *Models documentation*. https://platform.openai.com/docs/models/gpt-4 (visited on 07/21/2023)

OpenAI. (2023c). *Tokenizer*. https://platform.openai.com/tokenizer (visited on 12/20/2023)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. *35*, pp. 27730–27744). Curran Associates, Inc.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Parr, C. S., Wilson, N., Leary, P., Schulz, K. S., Lans, K., Walley, L., Hammock, J. A., Goddard, A., Rice, J., Studer, M., Holmes, J. T. G., & Corrigan, R. J., Jr. (2014). The encyclopedia of life v2: Providing global access to knowledge about life on earth. *Biodiversity Data Journal*, *2*, e1079. https://doi.org/10.3897/BDJ.2.e1079

Peng, A., Wu, M., Allard, J., Kilpatrick, L., & Heidel, S. (2023). *GPT-3.5 Turbo finetuning and API updates*. https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates (visited on 01/03/2024)

Scheepens, D. (2024). Dscheepens/GPT4-pest-controllers: Public release. Version v1.2 https://doi.org/10.5281/zenodo.10969983

Silva, A. D., Wijekoon, J. L., Liyanarachchi, R., Panchendrarajan, R., & Rajapaksha, W. (2024). *AI insights: A case study on utilizing ChatGPT intelligence for research paper analysis*.

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., … Hartshorn, A. (2023). *Llama 2: Open foundation and fine-tuned chat models*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention is all you need*.

von Dohlen, C. D., & Moran, N. A. (1995). Molecular phylogeny of the homoptera: A paraphyletic taxon. *Journal of Molecular Evolution*, *41*(2), 211–223. https://doi.org/10.1007/BF00170675

Wagner, D. L., Grames, E. M., Forister, M. L., Berenbaum, M. R., & Stopak, D. (2021). Insect decline in the Anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(2), e2023989118. https://doi.org/10.1073/pnas.2023989118

Wang, H., Fu, T., du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T. Y., Manrai, A., … Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9. https://doi.org/10.1186/s40537-016-0043-6

Wratten, S. D., Gillespie, M., Decourtye, A., Mader, E., & Desneux, N. (2012). Pollinator habitat enhancement: Benefits to other ecosystem services. *Agriculture, Ecosystems & Environment*, *159*, 112–122. https://doi.org/10.1016/j.agee.2012.06.020

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). *BloombergGPT: A large language model for finance*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). *OPT: Open pre-trained transformer language models*.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2023). *Least-to-Most prompting enables complex reasoning in large language models*.

SUPPORTING INFORMATION

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

Table S1. Glossary of key terms.

Table S2. Analysis of mismatches between manual labels and GPT-4 predictions of taxonomic terms available in the abstract (a) and missing from the abstract (b).

Table S3. Proportion correct (PC) scores of GPT-4 on (a) taxonomic named entity recognition and (b) higher level taxonomy prediction.

Table S4. Precision, recall and F1-score obtained by GPT-4 for each role in the training set (a) and the held-out test set (b), computed as the mean (with a spread of one standard deviation) over all abstracts in which the role is either manually labelled or predicted by GPT-4.

Table S5. Precision and recall obtained by GPT-4 for geographic locations in the training set (a) and the held-out test set (b), computed as the mean (with a spread of one standard deviation) over all abstracts in which the label (grouped as 'Location' and 'No location') is either manually labelled or predicted by GPT-4.

Table S6. Precision and recall obtained by GPT-4 for the pest industry types (column 11) in the training set (a) and the test set (b).

Table S7. Confabulation check for pest industries in the training set.

Table S8. Confabulation check for pest industries in the test set.

Figure S1. Occurrences (mention of a single term in a single abstract) of taxonomic orders and ascribed roles in the (manually labelled) training set (a) and the held-out test set (b). Both data sets comprise 100 abstracts.

Figure S2. Table generated by GPT-4 for an exemplar abstract (title, abstract and keywords) in the test set, using the fine-tuned prompt design (Fig. 1).

Figure S3. Table generated by GPT-4 for an exemplar abstract (title, abstract and keywords) in the test set, using the fine-tuned prompt design.

Figure S4. Confusion matrix of the various roles ascribed to species, either as 'Manual label' or as 'GPT-4 prediction', for the training set.

Figure S5. Confusion matrices of the geographic locations of the studies mentioned in the text, either as 'Manual label' or as 'GPT-4 prediction', for the training set (a) and the held-out test set (b).

Figure S6. Confusion matrices of the industry types of pests and pest controllers, either as 'Manual label' or as 'GPT-4 prediction', for the training set (a) and the held-out test set (b).

Figure S7. Prompt version 1: Initial prompt design.

Figure S8. Prompt version 2: Prompt design following the first iteration of fine-tuning against the training set.

Figure S9. Prompt version 3: Prompt design attempt during the second iteration of fine-tuning against the training set.

Figure S10. Prompt version 4: Prompt design attempt during the second iteration of finetuning against the training set.

Figure S11. Prompt version 5: Prompt design following the second iteration of fine-tuning against the training set.

Figure S12. Prompt version 6: Prompt design attempt during the third iteration of fine-tuning against the training set.

Figure S13. Prompt version 7: The final prompt design, following the third iteration of finetuning against the training set.

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

2041210x, 2024, 7, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14341 by <Shibboleth>-member@imperial.ac.uk, Wiley Online Library on [22/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License