bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Extracting massive ecological data on state and interactions of species using large language models

François Keck[1,2], Henry Broadbent[1,2], Florian Altermatt[1,2]

[1] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland

[2] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department of Aquatic Ecology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Abstract

The contemporary ecological crisis calls for integration and synthesis of ecological data describing the state, change and processes of ecological communities. However, such synthesis depends on the integration of vast amounts of mostly scattered and often hard-to-extract information that is published and dispersed across hundreds of thousands of scientific papers, for example describing species-specific interactions and trophic relationships. Recent advancements in natural language processing (NLP) and in particular the emergence of large language models (LLMs) offer a novel, and potentially revolutionary solution to this persistent challenge, for the first time creating the opportunity to access and extract virtually all data ever published. Here, we demonstrate the transformative potential of LLMs by extracting all types of biological interactions among species directly from a corpus of 83,910 scientific articles. Our approach successfully extracted a network of 144,402 interactions between 36,471 taxa. Performance analysis shows that the model exhibits a high sensitivity (70.0%) and excellent precision (89.5%). Our approach proves that LLMs are capable of carrying out complex extraction tasks on key ecological data on a very large scale, paving the way for a multitude of potential applications in ecology and beyond.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Main

As stated by Rachel Carson “In nature, nothing exists alone”, it is not the sheer number of species, but rather their interactions and dependencies that define ecological systems. These biological interactions, such as predation, competition, parasitism or mutualism, are the hallmark of ecology and biodiversity sciences. They are a key concept in ecology, and form the foundation of population dynamics, community composition and ecosystem functioning[1–3]. Biological interactions, together with species distribution and functional traits, also represent the key critical information for understanding and responding to increasing environmental challenges[4] including climate change[5], biodiversity loss[6,7], and emerging pathogens[8].

Yet, data on biological interactions are among the hardest to get hold on. Information on the various types on how species interact is largely based on original natural history observations scattered across a myriad of publications and often dispersed across different research fields. Also, the possible number of interactions is vastly larger than the sum of species involved. Imagine a community of 200 plants and 600 insect herbivores–realistic numbers for even a mid-diversity ecosystem–creating the potential of close to 20,000 pairwise competitive interactions between the plants and close to 60,000 herbivorous interactions alone. Despite only a very sparse part of these potential interactions being realized, assembling and integrating data on species interactions has been a major challenge, with only few food-webs being largely resolved[9,10], yet alone thinking beyond pairwise interactions[11].

Because of the central importance of biological interactions for modern ecology and for biodiversity prediction in the current context of global change[11–15], large database projects have been developed with the aim of collecting machine-readable species-interaction data[16–18]. However, these databases ultimately rely on manual or semi-automatic extraction and integration of data shared by scientists. While data on species occurrences are rapidly

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

increasing and are expected to become less of a limiting factor for ecological research[19], information on species interactions remains significantly lacking and incomplete[20–22]. This discrepancy arises partly due to the inherent complexity of sampling species interactions[23] and partly because such data are less systematically recorded and reported.

Integrating and synthesizing ecological knowledge is notoriously difficult. This is largely due to how ecology and natural history findings are disseminated. Typically, they are communicated through scientific articles, which often consist of large volumes of unstructured text[24]. Consequently, the sum of ecological knowledge generated over more than a century, and which we need to mobilize immediately to tackle the global ecological crisis, is dissolved in a vast ocean of text. Unlike structured databases or standardized formats, scientific articles vary widely in their organization. This has made extraction of information laborious and time-consuming or even impossible, such that most information is still hidden (if not lost) in text. Also, ecological research spans diverse subfields and methodologies, resulting in a large and heterogeneous body of literature that can be difficult to navigate and integrate. As a result, ecologists face the daunting task of sifting through a myriad of articles, extracting relevant data, and synthesizing findings into cohesive reviews, databases and meta-analyses. Cataloguing interactions among species started with the very first steps of ecology by Alexander von Humboldt and the tangled bank by Darwin, yet despite generations of ecologists and naturalists collecting and documenting these interactions, they could have been hardly integrated in an inclusive manner.

The current revolution in text mining and natural language processing (NLP) driven by the development of deep neural approaches[25] and more recently by the advent of large language models (LLMs) is for the first time opening the potential for automated data extraction and synthesis from scientific articles. These models trained on vast amounts of text data, now exhibit unprecedented capabilities in understanding and generating human-like language, allowing to tackle more complex extraction tasks which until now were limited to highly structured and codified textual data (e.g. collections of taxonomic

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

descriptions[26]). In academia the groundbreaking potential of LLMs, including ChatGPT, the popular LLM-powered chatbot by OpenAI, have predominantly been explored and discussed in the context of text generation (e.g. for writing publications[27]) and education (e.g. adoption and use by students and teachers[28–30]), while the capacity of LLMs for information extraction in the form of structured data[31,32] has received comparatively little attention. First case studies in ecology and biodiversity science use LLMs for automatic data extraction from disease reports[33], research abstracts[34,35] and news reports[35]. Nevertheless, these works remain proof-of-concept conducted on small and therefore limited[36] samples of text, and focusing on specific issues or taxa. Large-scale implementation for more general problems and on a corpus representative of the state of knowledge remains to prove that LLMs can truly revolutionize the synthesis and analysis of ecological knowledge.

Here, we demonstrate the power of LLMs to extract biotic interactions between species directly from the published literature at a very large scale. As shown above, biological interactions are a critically important piece of information for ecological research, and their availability is still very limited. But beyond the hope raised by LLMs to solve this long standing problem, biological interactions are an interesting case study, as they represent a technical challenge for automatic data extraction methods. This is complex information involving two actors (two species) and a directional relationship (the interaction). The algorithm or model has to isolate and qualify these elements, despite the extensive vocabulary and sometimes ambiguous terminology[37,38] that covers the wide range of biological interactions and the organisms involved. While dictionary-based approaches have been developed to detect biological interactions[39], their efficiency is strongly limited when confronted with the complexity of the language and the multiplicity of syntactic constructions used in the literature. To date, this complexity precluded any attempt at automated extraction of biological interactions on a large scale.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Results

From 545,967 processed paragraphs originating from 83,910 scientific articles, the model extracted 649,319 potential biotic interactions. These interactions involved 157,354 unique named entities (theoretically representing names of organisms) among which 70,841 could be taxonomically linked to the NCBI Taxonomy database through TEL. We excluded results from 29 paragraphs for which the model generated an infinite sequence of repeated words and also 4,272 interactions that matched exactly the examples provided in the prompt and were thus most likely the result of an overfit. To merge synonymous nodes and remove potential false positives, we filtered the data to keep only interactions involving taxonomically linked entities. We also merged all edges between two nodes belonging to the same category. This resulted in a final interaction network of 35,471 nodes and 144,402 edges (interactions) originating from 112,647 paragraphs, from 36,044 publications. The interactions were described in the source publications with a very rich and more or less specific vocabulary (14,767 unique labels, see Fig. 2 for an overview of label-to-category associations), demonstrating the importance of a good understanding of the context by the model.

The global set of extracted interactions were filtered and aggregated at different levels of taxonomic resolution (Table 1). For example, the species level network involved 18,589 different species and 46,467 species-to-species interactions distributed in 1,295 components (see Fig. 1 for a detailed view of this network). Interactions were dominated by parasitism (Fig. 1a-c) and the species with the most interactions extracted were species exhibiting a major role in human cultures, such as domestic animals and livestock, or parasites of medical or veterinary interest (Fig. 1d).

To compare and validate these LLMs extracted interactions, we manually screened a subset of this text and extracted all interactions based on expert knowledge. The manual

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

annotation of 500 paragraphs in the validation set identified 327 biological interactions. Our automated approach managed to identify 229 of these (true positives), while making 27 errors (false positives). This corresponds to an accuracy of 89.5% and a recall of 70.0%.

Discussion

As environmental challenges are multiplying and becoming more pressing, it becomes increasingly important to mobilize all available knowledge in ecology. Here, we successfully demonstrated that this is now within reach and made possible by new developments in artificial intelligence and, in particular, the emergence of high-performance large language models. We extracted detailed information on biological interactions, a key parameter in ecology, among thousands of species directly from the scientific literature and at unprecedented scales, with the opportunities to further integrate these data into existing products or to produce original ecological research about the structure and properties of this global ecological meta-network[40,41]. Up to now, access to species interaction data has been one of the most limited and biased, yet one of the most sought after data types[20,22]. These results give us a glimpse of a future where the literature will be scanned on large scales to extract massive amounts of data that were previously beyond the reach of any automated algorithm.

Our study also provides an opportunity to measure the accuracy of automatic LLM extraction in an ambitious and complex case study, and to identify the challenges and limitations of this novel approach when applied at large scale. Our method effectively balanced overall correctness with the ability to identify relevant cases, even in challenging scenarios. While the recall (70.0%) suggests room for improvement in detecting all relevant interactions, the relative high accuracy (89.5%) highlights the robustness of the model in minimizing false positives. Although the extraction relies primarily on a LLM, it is part of a more complex data processing pipeline that guarantees the quality of results. In particular,

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

we implemented a pre-filtering step to target the most relevant documents and a post-filtering step to improve the quality of the results. Pre-filtering is not strictly necessary but allows to restrict the corpus of text, hence limiting the number of requests and therefore reducing the financial and environmental costs of the model. In our experiment, post-filtering proved to be necessary to remove the large number of false positives extracted by the model (e.g. non relevant relationships between an organism and its environment). All these steps, when properly implemented, can be easily automated.

Importantly, as vast and rich as the scientific literature may be, it is inevitably biased, and the data extracted by our approach precisely reflects these biases. The species most represented in our results include humans (*Homo sapiens*), human pathogens (e.g. *Escherischia coli*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*) and domestic species (e.g. *Canis lupus*, *Sus scrofa*, *Bos taurus*), all of which have been the focus of particular attention from the scientific community. Data extraction can only highlight biases in the literature and does not correct them. This is a factor that must always be considered when examining the results.

A number of challenges and limitations have also been identified regarding the use of proprietary LLMs for scientific research[42]. First, LLMs can pose a problem for reproducibility because outputs can be non-deterministic, i.e., the same query (prompt) might yield different results depending on the model's internal state. A potential solution is to perform deterministic sampling by using fixed random seeds. This approach makes outputs more consistent across runs but all conditions (hardware, software, model parameters) must be controlled, which is difficult to ensure with cloud-based closed models. Secondly, these models are essentially black boxes. The data and methodology used to train them are generally kept confidential by the developing companies. Such opacity requires researchers to take a step back from the possible biases induced by the methodological choices made by the model's designers. Analysing the model's outputs carefully and performing a

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

thorough validation, as done in the present study, is therefore indispensable for any scientific application of this type of approach.

The above challenges are solvable, at least in part, through the use of small open models that are smaller in size but can be optimized and used locally. Despite their smaller size, small-scale LLMs – when appropriately fine-tuned – can achieve impressive results across different applications[43–45]. The advantages of this approach are many[46]: greater independence, greater transparency, better control on reproducibility, and potentially lower costs and environmental imprint. Moreover, the use of a model run locally potentially enables processing documents with more restrictive licenses[47]. However, selection, deployment, optimisation and fine-tuning of such models may require a certain amount of expertise and resources. Despite these challenges, it is clear that the impact of LLMs will be all the greater if they are put directly into the hands of scientists and make their way into laboratories[46,48]. In this respect, small-scale fine-tuned LLM represents a promising direction.

Our approach represents a potentially major paradigm shift with regard to the way we interact with scientific literature, fitting in with the epistemological revolution initiated by the emergence of LLMs[49]. Through our results we show that a new level of qualitative and quantitative synthesis of the available knowledge can be achieved. Building on recently published proof-of-concepts[33–35], our study grounds the use of LLMs for synthesis sciences and by demonstrating that we can extract accurate structured information from scientific publications on a large scale, it opens the way to a virtually infinite number of applications. Whether for building databases or conducting meta-analyses, the number of questions that can be addressed in ecology is significant. Automated data extraction also paves the way for living syntheses, i.e., studies that are continually updated and whose results and conclusions evolve as new primary results are made available[50]. In the context of the ongoing biodiversity crisis where scientists are confronted with an inflation in the number of publications and the dispersal of information, advancing ecological data access and

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

extraction to a new level will help to better understand and model current and future species distribution and interactions and guide critical management decisions.

Methods

Corpus compilation

We screened scientific publications using a large language model to extract all possible biological interactions. The text corpus was sourced from the PubMed Central (PMC) Open Access database, which contains millions of publications spanning various scientific disciplines[51]. To narrow the scope of our analysis, we first refined our corpus to focus specifically on publications related to ecological sciences. Using the OpenAlex database[52] and its topic classification system, we identified 99,555 publications tagged under the subfield "Ecology" from an initial pool of 6,159,719 PMC publications. We downloaded the corresponding 99,555 XML files from PMC and parsed them to extract 3,552,030 paragraphs for further analysis.

We refined this dataset by retaining only the paragraphs that included at least two distinct taxonomic names and one keyword associated with species interactions. To identify taxonomic names, we employed TaxoNERD[53], a named-entity recognition (NER) tool utilizing deep neural models for recognizing both scientific and vernacular taxonomic names. TaxoNERD detected a total of 8,160,227 taxonomic names across 1,925,414 paragraphs from 97,681 articles. In parallel, we performed a regular expression-based search to detect species interaction terms using a curated list of 32 relevant keywords. The application of both filters—requiring a minimum of two distinct taxonomic names and one species interaction keyword—resulted in a final corpus consisting of 545,967 paragraphs from 83,910 articles.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Data extraction

The final dataset of 545,967 paragraphs was subsequently processed using the OpenAI GPT-4o model. Each paragraph was incorporated into a standardized prompt template (Fig. 3), which was then submitted to the OpenAI API for analysis. The prompt template was specifically designed to extract all pairwise species interactions present within the paragraph and structure the output in a tabular format. The output consisted of a four-column table where the first two columns contained the names of the two species involved in the interaction, the third column specified the nature of the interaction as labeled in the text, and the fourth column classified the interaction into its respective category from the list used in the Mangal database[17]. This structured format allowed the systematic analysis and categorization of species interactions within the corpus.

To identify interactions involving entities that cannot be linked to living organisms and to harmonize organism names and link them to their upstream taxonomy, we performed taxonomic entity linking (TEL). TEL consists in mapping extracted named entities to corresponding unique identifiers in a target knowledge base, in our case the NCBI Taxonomy[54] . For example, through TEL, the entities dog, domestic dog, *Canis familiaris* and *Canis lupus familiaris* are all linked to the same NCBI Taxonomy ID: 9615 and constitute a unique node in the final network. We implemented TEL through a custom multistep algorithm using data from NCBI, the Encyclopedia of Life (EoL), the Integrated Taxonomic Information System (ITIS) and Wikipedia (see corresponding R script for detailed implementation).

Performance assessment

To estimate the quality and completeness of the information extracted by our approach, we manually reviewed and annotated 500 paragraphs (≈ 1% of the total corpus) taken at random from all those submitted to the model. Using these human annotated data, we

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

256 computed two performance statistics. First, the precision, calculated as the fraction of
257 relevant retrieved interactions among all the retrieved interactions and second, the recall,
258 calculated as the fraction of relevant interactions that were retrieved.

259 Data availability

260 The publications used to produce the results are all part of the PMC Open Access Subset
261 collection (https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/). All data extracted are openly
262 available at: https://github.com/fkeck/gpt_interactions

263 Code availability

264 All code used for analysis is available on GitHub: https://github.com/fkeck/gpt_interactions

265 References

266 1. Elton, C. S. *Animal Ecology*. 1–256 (Macmillan Co, New York, 1927).
267 doi:10.5962/bhl.title.7435.

268 2. Paine, R. T. Food Web Complexity and Species Diversity. *Am. Nat.* **100**, 65–75 (1966).
269 3. Seibold, S., Cadotte, M. W., MacIvor, J. S., Thorn, S. & Müller, J. The Necessity of
270 Multitrophic Approaches in Community Ecology. *Trends Ecol. Evol.* **33**, 754–764 (2018).
271 4. Pollock, L. J. *et al.* Protecting Biodiversity (in All Its Complexity): New Models and
272 Methods. *Trends Ecol. Evol.* **35**, 1119–1128 (2020).
273 5. Hegland, S. J., Nielsen, A., Lázaro, A., Bjerknes, A.-L. & Totland, Ø. How does climate
274 warming affect plant-pollinator interactions? *Ecol. Lett.* **12**, 184–195 (2009).
275 6. McDonald-Madden, E. *et al.* Using food-web theory to conserve ecosystems. *Nat.*
276 *Commun.* **7**, 10245 (2016).
277 7. Sandor, M. E., Elphick, C. S. & Tingley, M. W. Extinction of biotic interactions due to

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

habitat loss could accelerate the current biodiversity crisis. *Ecol. Appl.* **32**, e2608 (2022).

8. Wardeh, M., Baylis, M. & Blagrove, M. S. C. Predicting mammalian hosts in which novel coronaviruses can be generated. *Nat. Commun.* **12**, 780 (2021).
9. Pearse, I. S. & Altermatt, F. Extinction cascades partially estimate herbivore losses in a complete Lepidoptera–plant food web. *Ecology* **94**, 1785–1794 (2013).
10. Pearse, I. S. & Altermatt, F. Predicting novel trophic interactions in a non-native world. *Ecol. Lett.* **16**, 1088–1094 (2013).
11. Levine, J. M., Bascompte, J., Adler, P. B. & Allesina, S. Beyond pairwise mechanisms of species coexistence in complex communities. *Nature* **546**, 56–64 (2017).
12. Burkle, L. A. & Alarcón, R. The future of plant–pollinator diversity: Understanding interaction networks across time, space, and global change. *Am. J. Bot.* **98**, 528–538 (2011).
13. Wisz, M. S. *et al.* The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.* **88**, 15–30 (2013).
14. Staniczenko, P. P. A., Sivasubramaniam, P., Suttle, K. B. & Pearson, R. G. Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecol. Lett.* **20**, 693–707 (2017).
15. Cámara-Leret, R., Fortuna, M. A. & Bascompte, J. Indigenous knowledge networks in the face of global change. *Proc. Natl. Acad. Sci.* **116**, 9913–9918 (2019).
16. Poelen, J. H., Simons, J. D. & Mungall, C. J. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol. Inform.* **24**, 148–159 (2014).
17. Poisot, T. *et al.* mangal – making ecological network analysis simple. *Ecography* **39**, 384–390 (2016).
18. Maiorano, L., Montemaggiori, A., Ficetola, G. F., O'Connor, L. & Thuiller, W. TETRA-EU 1.0: A species-level trophic metaweb of European tetrapods. *Glob. Ecol. Biogeogr.* **29**,

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

1452–1457 (2020).

19. Feng, X. *et al.* A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. *Glob. Ecol. Biogeogr.* **31**, 1242–1260 (2022).

20. Hortal, J. *et al.* Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**, 523–549 (2015).

21. Cameron, E. K. *et al.* Uneven global distribution of food web studies under climate change. *Ecosphere* **10**, e02645 (2019).

22. Poisot, T. *et al.* Global knowledge gaps in species interaction networks data. *J. Biogeogr.* **48**, 1552–1563 (2021).

23. Chacoff, N. P. *et al.* Evaluating sampling completeness in a desert plant–pollinator network. *J. Anim. Ecol.* **81**, 190–200 (2012).

24. Poisot, T., Riva, G. D., Desjardins-Proulx, P. & Luccioni, A. S. The future of ecological research will not be (fully) automated. Preprint at https://doi.org/10.22541/au.169384322.27179185/v1 (2023).

25. Farrell, M. J. *et al.* The changing landscape of text mining: a review of approaches for ecology and evolution. *Proc. R. Soc. B Biol. Sci.* **291**, 20240423 (2024).

26. Coleman, D., Gallagher, R. V., Falster, D., Sauquet, H. & Wenk, E. A workflow to create trait databases from collections of textual taxonomic descriptions. *Ecol. Inform.* **78**, 102312 (2023).

27. Hutson, M. Could AI help you to write your next paper? *Nature* **611**, 192–193 (2022).

28. Kasneci, E. *et al.* ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).

29. Cooper, N., Clark, A. T., Lecomte, N., Qiao, H. & Ellison, A. M. Harnessing large language models for coding, teaching and inclusion to empower research in ecology and evolution. *Methods Ecol. Evol.* **15**, 1757–1763 (2024).

30. Campbell, H. *et al.* Should we still teach or learn coding? A postgraduate student perspective on the use of large language models for coding in ecology and evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Methods Ecol. Evol. **15**, 1767–1770 (2024).

31. Dagdelen, J. *et al.* Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
32. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **15**, 1569 (2024).
33. Gougherty, A. V. & Clipp, H. L. Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *Npj Biodivers.* **3**, 1–5 (2024).
34. Scheepens, D., Millard, J., Farrell, M. & Newbold, T. Large language models help facilitate the automated synthesis of information on potential pest controllers. *Methods Ecol. Evol.* **15**, 1261–1273 (2024).
35. Castro, A., Pinto, J., Reino, L., Pipek, P. & Capinha, C. Large language models overcome the challenges of unstructured text data in ecology. *Ecol. Inform.* **82**, 102742 (2024).
36. Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J. & Brunak, S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS Comput. Biol.* **14**, e1005962 (2018).
37. Hodges, K. E. Defining the problem: terminology and progress in ecology. *Front. Ecol. Environ.* **6**, 35–42 (2008).
38. Trombley, C. A. & Cottenie, K. Quantifying the Scientific Cost of Ambiguous Terminology in Community Ecology. *Philos. Top.* **47**, 203–218 (2019).
39. Muñoz, G., Kissling, W. D. & Loon, E. E. van. Biodiversity Observations Miner: A web application to unlock primary biodiversity data from published literature. *Biodivers. Data J.* **7**, e28737 (2019).
40. Delmas, E. *et al.* Analysing ecological networks of species interactions. *Biol. Rev.* **94**, 16–36 (2019).
41. Guimarães, P. R. The Structure of Ecological Networks Across Levels of Organization.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Annu. Rev. Ecol. Evol. Syst. **51**, 433–460 (2020).

42. Ollion, É., Shen, R., Macanovic, A. & Chatelain, A. The dangers of using proprietary LLMs for research. *Nat. Mach. Intell.* **6**, 4–5 (2024).

43. Alizadeh, M. *et al.* Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning. Preprint at https://doi.org/10.48550/arXiv.2307.02179 (2024).

44. Zhang, G. *et al.* Closing the gap between open source and commercial large language models for medical evidence summarization. *Npj Digit. Med.* **7**, 1–8 (2024).

45. Carammia, M., Iacus, S. M. & Porro, G. Rethinking Scale: The Efficacy of Fine-Tuned Open-Source LLMs in Large-Scale Reproducible Social Science Research. Preprint at https://doi.org/10.48550/arXiv.2411.00890 (2024).

46. Chen, L. & Varoquaux, G. What is the Role of Small Models in the LLM Era: A Survey. Preprint at https://doi.org/10.48550/arXiv.2409.06857 (2024).

47. Sommers, F., Kongthon, A. & Kongyoung, S. Fine-Tuning Large Language Models for Private Document Retrieval: A Tutorial. in *Proceedings of the 2024 International Conference on Multimedia Retrieval* 1319–1320 (Association for Computing Machinery, New York, NY, USA, 2024). doi:10.1145/3652583.3658419.

48. Ollion, E., Shen, R., Macanovic, A. & Chatelain, A. ChatGPT for Text Annotation? Mind the Hype! Preprint at https://doi.org/10.31235/osf.io/x58kn (2023).

49. Morera, A. Foundation models in shaping the future of ecology. *Ecol. Inform.* **80**, 102545 (2024).

50. Berger-Tal, O. *et al.* Leveraging AI to improve evidence synthesis in conservation. *Trends Ecol. Evol.* **39**, 548–557 (2024).

51. Roberts, R. J. PubMed Central: The GenBank of the published literature. *Proc. Natl. Acad. Sci.* **98**, 381–382 (2001).

52. Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. Preprint at https://doi.org/10.48550/arXiv.2205.01833 (2022).

53. Le Guillarme, N. & Thuiller, W. TaxoNERD: Deep neural models for the recognition of

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

taxonomic entities in the ecological and evolutionary literature. *Methods Ecol. Evol.* **13**, 625–641 (2022).

54. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2012).

Acknowledgments

Funding is by the University of Zurich “Fonds zur Förderung des Akademischen Nachwuchs” (to FK) and the Swiss National Science Foundation (grant 310030_197410) to FA.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Ethics declarations

Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Tables

Taxonomic level	Nodes	Edges	Components
Mixed (all)	36,471	144,402	933
Superkingdom	4	157	1
Phylum	127	5,201	2
Class	353	9,772	2
Order	1,140	26,345	11
Family	4,118	42,744	58
Genus	12,980	56,404	390
Species	18,589	46,467	1,295

Table 1. Summary statistics of the extracted network aggregated at different taxonomic levels. The nodes correspond to individual taxa, the edges represent the interactions between the taxa, and the components are the subparts of the main graph that are not connected together.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Figures

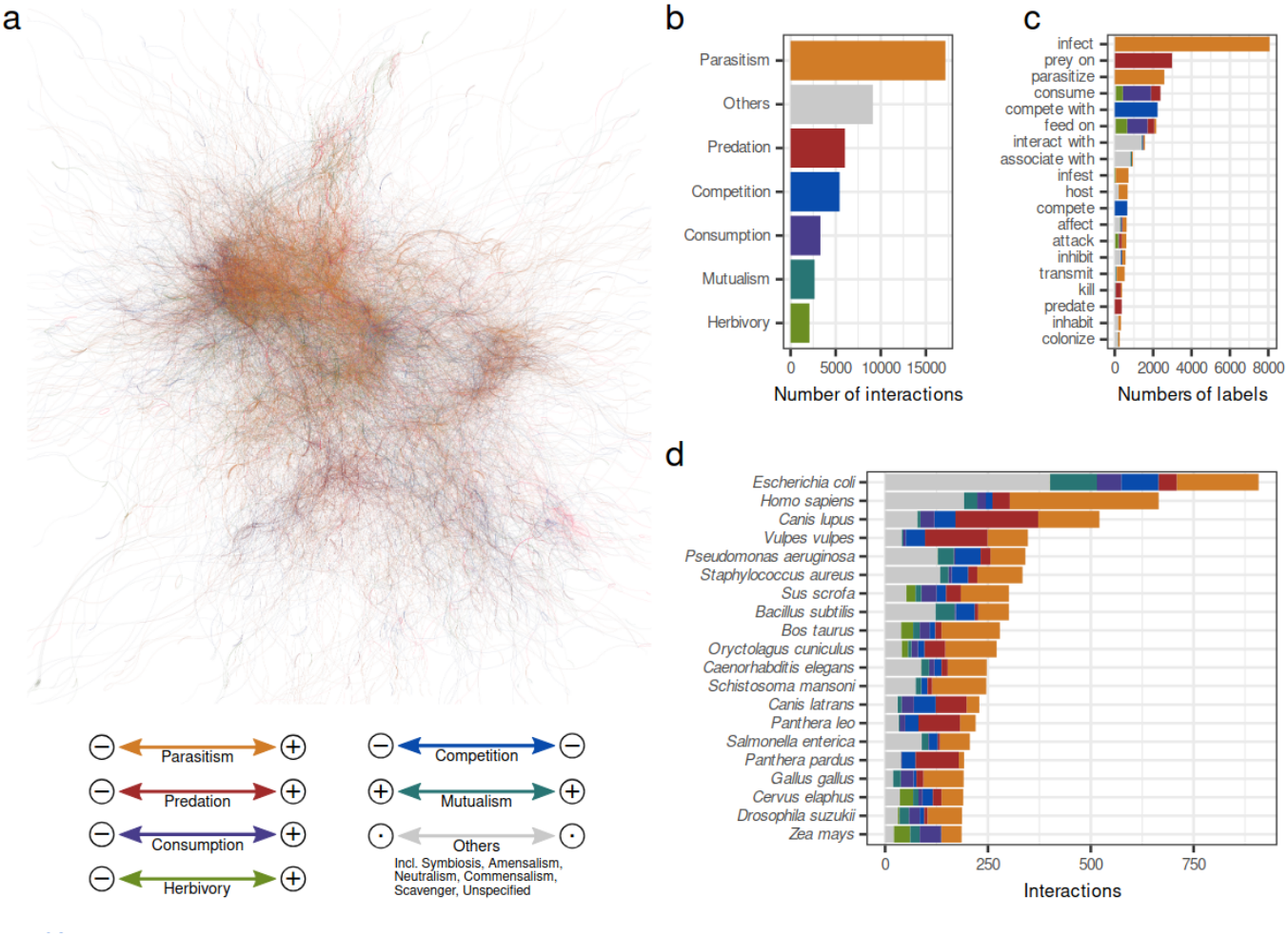


Figure 1. Reconstructed network of interactions at species level. **a.** Overview of the main graph component represented using a force-based (ForceAtlas2) layout. **b.** Number of interactions for the different inferred categories across the whole network. **c.** Number of extracted labels and their inferred categories for the 20 most common labels. **d.** Number of interactions and their inferred categories for the 20 most connected species. In each panel, interaction categories are color-coded as described in the legend, which also indicates for each category how the partners are affected (positively or negatively).

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

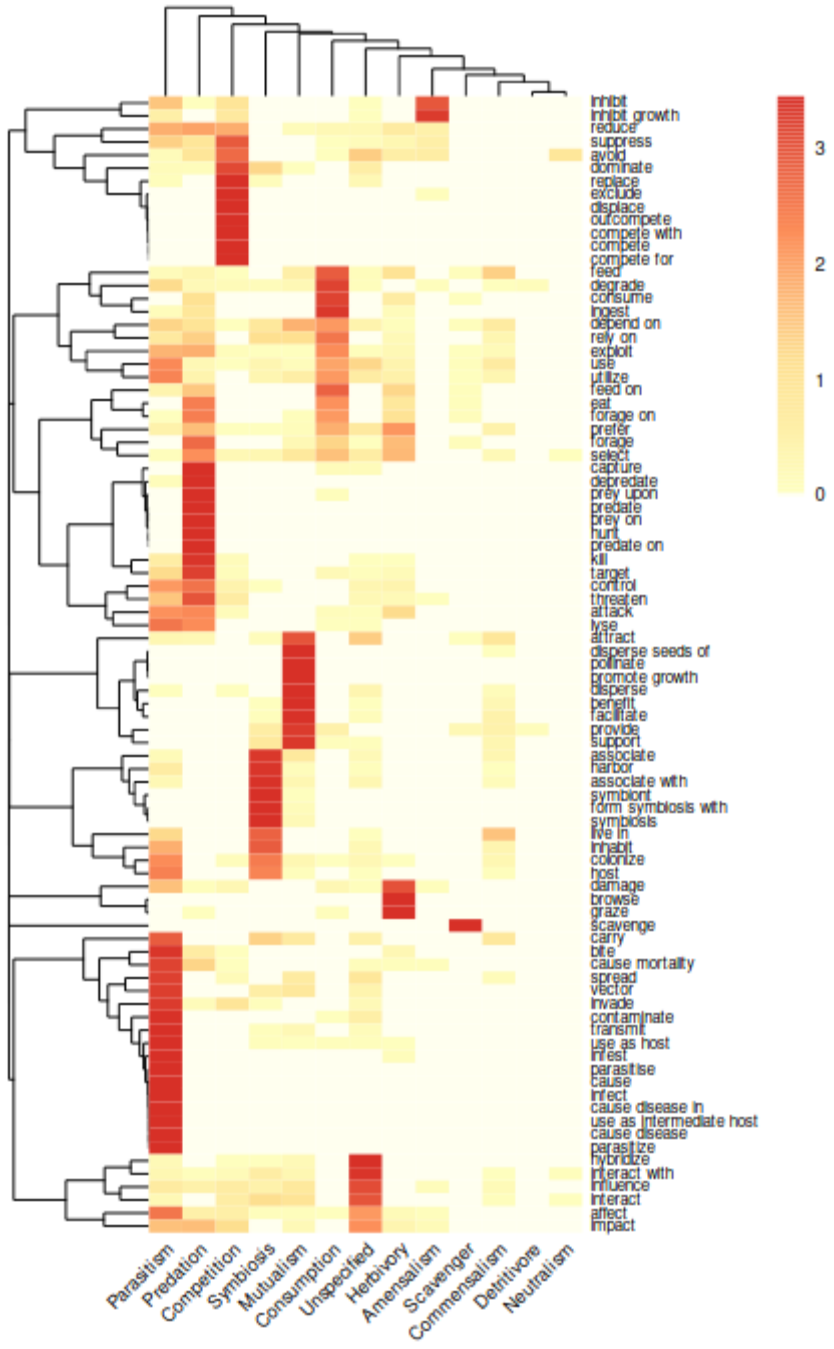


Figure 2. Heatmap showing the association between extracted labels and inferred interaction categories. Values represent the number of times each label has been classified into each category, scaled by row to improve readability. Only the most common labels (extracted more than 250 times) are shown (n = 87). The dendrograms represent the results of hierarchical clustering for rows (labels) and columns (categories).

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634685; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

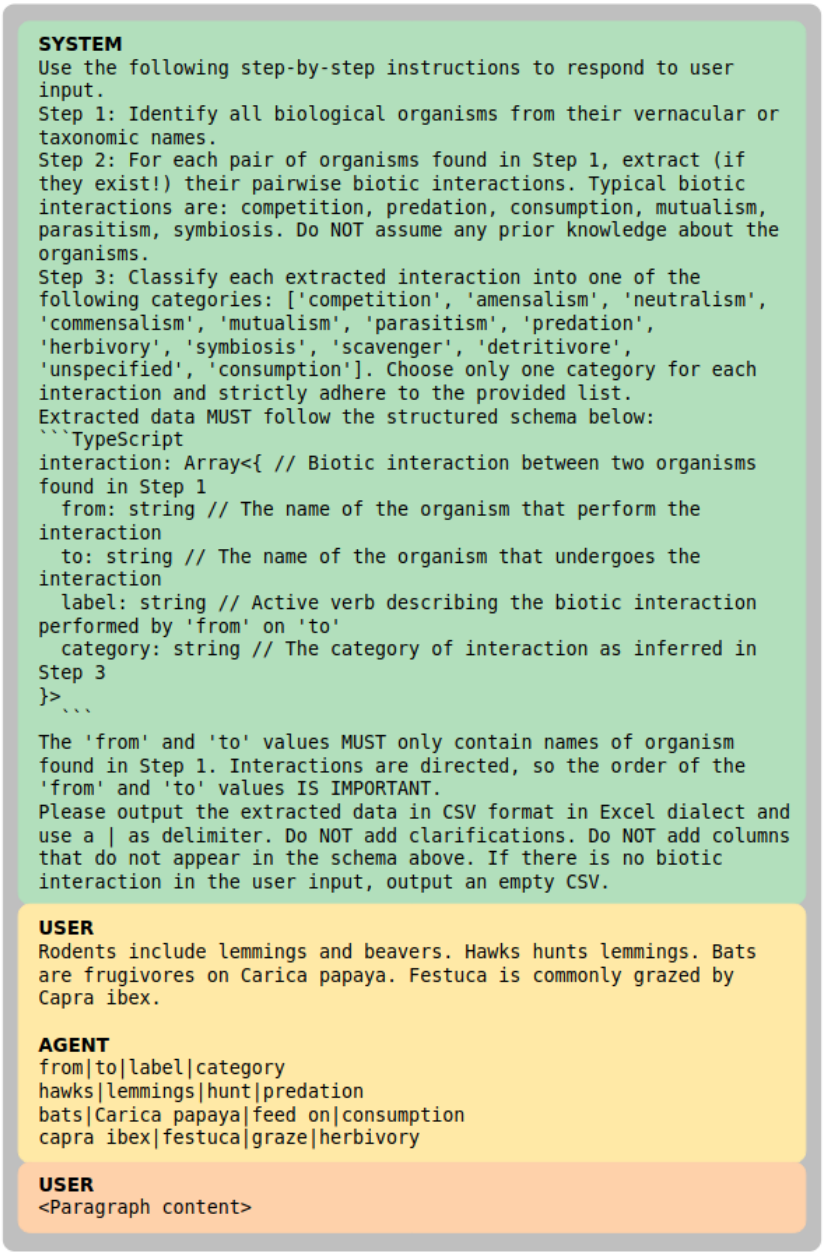


Figure 3. Standardised prompt template used to process each paragraph. First a system prompt is used to give a detailed description of the task to the model. Second, one example of task resolution is provided to the model to refine performances. Finally the content of the actual targeted paragraph is provided for extraction.