Automating Information Retrieval from Biodiversity Literature Using Large Language Models: A Case Study

Vamsi Krishna Kommineni[‡,§,|], Waqas Ahmed[‡], Birgitta Koenig-Ries[‡,¶,§], Sheeba Samuel[#,¶]

‡ Friedrich Schiller University Jena, Heinz Nixdorf Chair for Distributed Information Systems, Jena, Germany
§ German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
| Max Planck Institute for Biogeochemistry, Jena, Germany
¶ Michael Stifel Center Jena, Jena, Germany
# Chemnitz University of Technology, Distributed and Self-organizing Systems, Chemnitz, Germany

Corresponding author: Vamsi Krishna Kommineni (vamsi.krishna.kommineni@uni-jena.de)

Abstract

Recently, Large Language Models (LLMs) have transformed information retrieval, becoming widely adopted across various domains due to their ability to process extensive textual data and generate diverse insights. Biodiversity literature, with its broad range of topics, is no exception to this trend (Boyko et al. 2023, Castro et al. 2024). LLMs can help in information extraction and synthesis, text annotation and classification, and many other natural language processing tasks. We leverage LLMs to automate the information retrieval task from biodiversity publications, building upon data sourced from our previous work (Ahmed et al. 2024).

In our previous work (Ahmed et al. 2023, Ahmed et al. 2024), we assessed the reproducibility of deep learning (DL) methods used in biodiversity research. We developed a manual pipeline to extract key information on DL pipelines—dataset, source code, open-source frameworks, model architecture, hyperparameters, software and hardware specs, randomness, averaging result and evaluation metrics from 61 publications (Ahmed et al. 2024). While this allowed analysis, it required extensive manual effort by domain experts, limiting scalability. To address this, we propose an automatic information extraction pipeline using LLMs with the Retrieval Augmented Generation (RAG) technique. RAG combines the retrieval of relevant documents with the generative capabilities of LLMs to enhance the quality and relevance of the extracted information. We employed an open-source LLM, Hugging Face implementation of Mixtral 8x7B (Jiang et al. 2024), a mixture of expert models in our pipeline (Fig. 1) and adapted the RAG pipeline from earlier work (Kommineni et al. 2024). The pipeline was run on a single NVIDIA A100 40GB graphics processing unit with 4-bit quantization.

To evaluate our pipeline, we compared the expert-assisted manual approach with the LLM-assisted automatic approach. We measured their consistency using the inter-

annotator agreement (IAA) and quantified it with the Cohen Kappa score (Pedregosa et al. 2011), where a higher score indicates more reliable and aligned outputs (1: maximum agreement, -1: no agreement). The Kappa score among human experts (annotators 1 and 2) was 0.54 (moderate agreement), while the scores comparing human experts with the LLM were 0.16 and 0.12 (slight agreement). The difference is partly due to human annotators having access to more information (including code, dataset, figures, tables and supplementary materials) than the LLM, which was restricted to the text itself. Given these restrictions, the results are promising but also show the potential to improve them by adding further modalities to the LLM inputs.

Future work will involve several key improvements to our LLM-assisted information retrieval pipeline:

1. Incorporating multimodal data (e.g., figures, tables, code, etc.) as input to the LLM, alongside text, to enhance the accuracy and comprehensiveness of the information retrieved from publications.

2. Optimizing the retrieval component of the RAG framework with advanced techniques like semantic search, hybrid search or relevance feedback can improve the quality of outputs.

3. Expanding the evaluation to a larger corpus of biodiversity literature could provide a more comprehensive understanding of pipeline capabilities, and this paves the way for pipeline optimization.
4. A human-in-the-loop approach for evaluating the LLM-generated outputs by matching the ground truth values from the respective publications, will increase the quality of the overall pipeline.
5. Employing more metrics for the evaluation beyond the Cohen Kappa score to better understand the LLM-assisted outputs.

Leveraging LLMs to automate information retrieval from biodiversity publications signifies a notable advancement in the scalable and efficient analysis of biodiversity literature. Initial results show promise, yet there is substantial potential for enhancement through the integration of multimodal data, optimized retrieval mechanisms, and comprehensive evaluation. By addressing these areas, we aim to improve the accuracy and utility of our pipeline, ultimately enabling broader and more in-depth analysis of biodiversity literature.

Keywords

Large Language Models (LLMs), information retrieval, deep learning, Retrieval Augmented Generation (RAG), biodiversity

Presenting author

Vamsi Krishna Kommineni

Presented at

SPNHC-TDWG 2024

Acknowledgements

We acknowledge computing time on the HPC cluster Draco provided by the IT centre of the Thuringian universities.

Funding program

Supported by the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, funded by the German Research Foundation (FZT 118, 202548816) and Carl Zeiss Foundation for the project “A Virtual Werkstatt for Digitization in the Sciences (K3)” within the scope of the program line “Breakthroughs: Exploring Intelligent Systems for Digitization-Explore the Basics, Use Applications.”

Hosting institution

Friedrich Schiller University Jena, Jena, Germany

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Ahmed W, Kommineni VK, Koenig-ries B, Samuel S, et al. (2023) How Reproducible are the Results Gained with the Help of Deep Learning Methods in Biodiversity Research? Biodiversity Information Science and Standards 7 https://doi.org/10.3897/biss.7.112698
- Ahmed W, Kommineni VK, König-Ries B, Gaikwad J, Gadelha L, Samuel S, et al. (2024) Evaluating the method reproducibility of deep learning models in the biodiversity domain. arXiv https://doi.org/10.48550/arxiv.2407.07550
- Boyko J, Cohen J, Fox N, Veiga MH, Li JI, Liu J, Modenesi B, Rauch A, Reid K, Tribedi S, Visheratina A, Xie X, et al. (2023) An Interdisciplinary Outlook on Large Language Models for Scientific Research. arXiv https://doi.org/10.48550/arxiv.2311.04929
- Castro A, Pinto J, Reino L, Pipek P, Capinha C, et al. (2024) Large language models overcome the challenges of unstructured text data in ecology. bioRxiv https://doi.org/10.1101/2024.01.23.576654
- Jiang A, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, Chaplot DS, Casas Ddl, Hanna EB, Bressand F, Lengyel G, Bour G, Lample G, Lavaud LR, Saulnier L,

Lachaux M, Stock P, Subramanian S, Yang S, Antoniak S, Scao TL, Gervet T, Lavril T, Wang T, Lacroix T, Sayed WE, et al. (2024) Mixtral of Experts. arXiv https://doi.org/10.48550/arxiv.2401.04088

- Kommineni VK, König-Ries B, Samuel S (2024) From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. arXiv https://doi.org/10.48550/arxiv.2403.08345
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (85): 2825-2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html

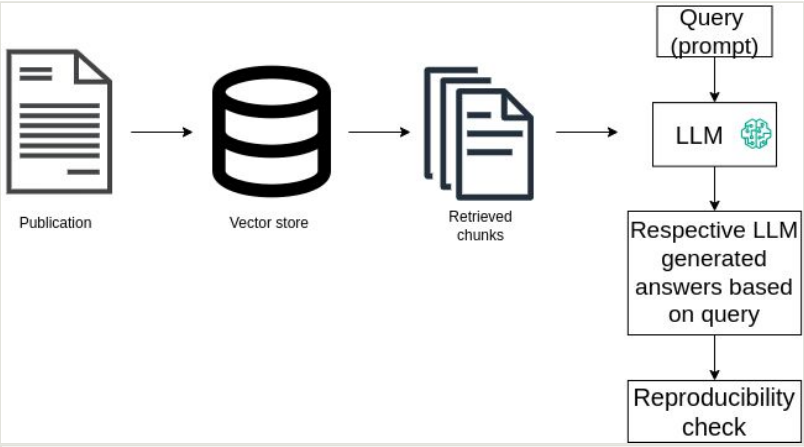

Figure 1.

Basic workflow of the RAG approach used in this study.