

An AI-powered knowledge base for scientific abstracts: a case study on environmental DNA (eDNA) in biomonitoring

Vesselin Baev^{1,*}, Gana Gecheva^{1,2}

¹Faculty of Biology, University of Plovdiv "Paisii Hilendarski", 24 Tsar Asen Str., Plovdiv, 4000, BULGARIA

^{1,2}Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Yuri Gagarin 2, 1113 Sofia, BULGARIA

*Corresponding author: baev@uni-plovdiv.bg

Abstract. Environmental DNA (eDNA) refers to genetic material shed by organisms into their environment, such as water, soil, or air. As a non-invasive biomonitoring method, eDNA has revolutionized biodiversity assessment by enabling the detection of species presence without direct observation or capture. This approach is especially critical for tracking invasive, elusive, or endangered species and monitoring ecosystem changes due to climate or anthropogenic pressures. Over the past decade, a growing body of scientific literature has explored eDNA applications, resulting in a fragmented but rich landscape of domain-specific knowledge. Navigating this information is increasingly challenging for researchers and policymakers. To address this, we developed BioTrace, an AI-powered knowledge base designed to support conversational exploration of scientific abstracts focused on eDNA in biodiversity monitoring. BioTrace leverages a Retrieval-Augmented Generation (RAG) architecture, integrating the mistral-saba-24b large language model via the Groq API for ultra-fast, low-latency inference. Scientific abstracts are indexed using a vector store, and retrieved passages are reranked using the all-MiniLM-L6-v2 model to improve answer relevance. Users can query the system in natural language and receive grounded, context-aware responses that synthesize findings across multiple studies. So far, the knowledge base includes more than 4000 abstracts on eDNA studies. This work demonstrates the potential of large language models (LLMs) to distill scientific literature into accessible, structured knowledge. BioTrace empowers users with real-time, interpretable insights into eDNA research, serving as a blueprint for future AI-based tools in ecological and environmental sciences.

Key words: AI, LLM models, RAG, eDNA, biomonitoring.

Introduction

Artificial intelligence (AI) represents a promising avenue for supporting researchers in their work. In particular, large language models (LLMs) such as the GPT family are advanced AI systems capable of understanding and generating human-like text based on extensive training data (Thirunavukarasu et al., 2023). These models can process vast amounts of literature much faster than manual methods, potentially reducing the time required for researchers to review publications. However, in some cases of outputs still demand

careful human verification to ensure accuracy and reliability. One of the key limitations of LLMs lies in their restricted context window, which constrains the volume of text they can process at once (Shi et al., 2023). This limitation poses challenges when analyzing lengthy scientific documents, such as those encountered in systematic reviews.

To address this issue, researchers have introduced the retrieval-augmented generation (RAG) framework (Lewis et al., 2020). RAG integrates retrieval-based approaches with generative models, allowing LLMs to access and utilize infor-

mation beyond their built-in context window. In this framework, a retrieval component identifies the most relevant sections from a large corpus using vector-based search methods, which are then provided to the LLM for response generation. This enables the model to draw on a broader informational base, enhancing its performance on complex tasks. RAG's ability to manage large volumes of text makes it particularly suitable for systematic reviews (SRs), improving processes such as article identification, selection, and summarization. Consequently, it can help streamline SR workflows while maintaining methodological rigor.

A major portion of the SR workload involves title and abstract screening. This step, which typically follows the formulation of a search strategy and precedes full-text screening, requires researchers to assess whether a publication should be included in the review based on its title and abstract. Automating this stage has the potential to significantly reduce workload, accelerate evidence synthesis, and improve methodological consistency and reproducibility. As such, the automation or semi-automation of title and abstract screening has long been an area of active research (Clark et al., 2020; Pham et al., 2021; van de Schoot et al., 2021).

LLMs have demonstrated impressive capabilities across diverse domains and tasks that typically demand deep comprehension and contextual reasoning from human experts. Their applications include classification (McNichols et al., 2023), information extraction (Wadhwa et al., 2023), and knowledge retrieval (Trajanoska et al., 2023). Moreover, they can be flexibly adapted through prompt engineering (Reynolds & McDonnell, 2021) and parameter tuning to perform specific roles. Nonetheless, challenges remain, such as model "hallucinations" (Guerreiro et al., 2023), inherent biases (Hastings, 2024; Zack et al., 2024), and imperfect alignment with human judgment (Digtusch & Kosinski, 2023). Despite these issues, employing LLMs for evaluating scientific publications represents an innovative and promising approach toward developing automated yet adaptable systems for screening and assessing scientific literature.

Recent advancements in the field of environmental DNA (eDNA) have been remarkable (Taberlet et al., 2018). Although the core principles and workflows of DNA barcoding and metabar-

coding remain largely unchanged, rapid progress in high-throughput sequencing technologies has significantly improved access to eDNA data. These innovations have enabled biodiversity monitoring with unprecedented precision and scale, generating massive environmental genomic datasets at relatively low cost. The application of machine learning and other taxonomy-free analytical methods has opened new opportunities to identify bioindicators and assess ecological status (Cordier et al., 2018, 2019; Pawlowski et al., 2018). As scientific literature in this field continues to expand rapidly, LLMs can provide substantial support to researchers by facilitating data interpretation and literature synthesis.

This study demonstrates the potential of large language models to transform complex scientific abstract information into structured, accessible knowledge. We developed BioTrace, a tool designed to offer researchers real-time, transparent insights into eDNA research, serving as a model for the next generation of AI-driven tools in ecology and environmental science.

Materials and methods

Abstract source

Abstracts were fetched in bulk from the Scopus database (<https://www.scopus.com/>).

System overview

The tool is written in Python, and demo version is implemented at HuggingFace - an open-source artificial intelligence (AI) and machine learning (ML) platform (<https://huggingface.co/>).

Retriever and Reranking layer

This module employs the cross-encoder/ms-marco-MiniLM-L-6-v2 model. The model ms-marco-MiniLM-L-6-v2 is a sentence embedding model created by Microsoft and available via the Sentence Transformers library.

Conversational querying and LLM model

The conversational reasoning component was powered by model llama-3.1-8b-instant, an LLM optimized for rapid inference.

Data visibility

The source code can be accessed via <https://huggingface.co/spaces/vebaev/BioTrace/blob/main/app.py>

Results and Discussion

We developed BioTrace, a tool for AI-assisted analysis of scientific abstracts, focusing on eDNA studies. The system implements a retrieval-augmented generation (RAG) workflow, combining

vector-based document retrieval, cross-encoder-based reranking, and large language model (LLM) reasoning to enable interactive querying of Scientific knowledge.

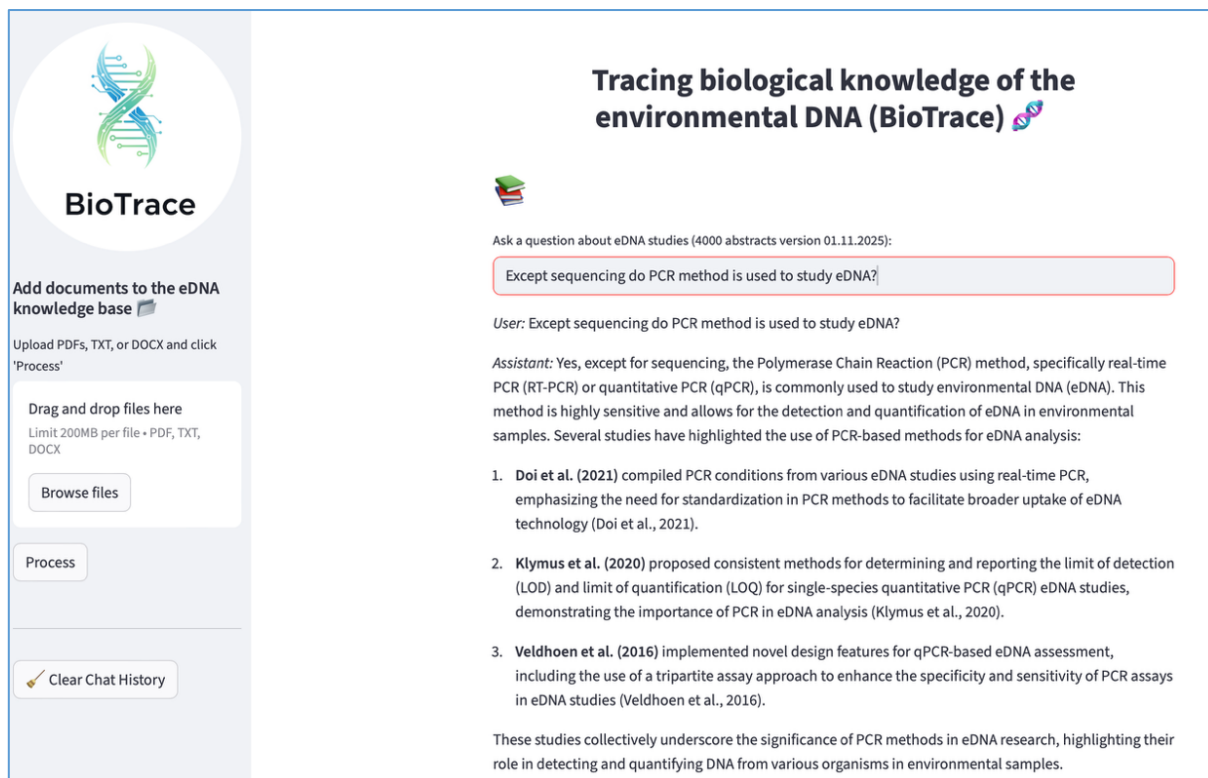


Fig 1. User interface of the BioTrace tool.

The user interface (UI) of the BioTrace tool is designed with clarity and usability in mind, aiming to streamline the process of querying and expanding the eDNA knowledge base of scientific abstracts (Fig. 1). Users can easily drag and drop new documents to contribute to the database, with feedback on file limits and types clearly displayed. Central to the UI is a responsive query input field where users can submit questions about eDNA studies. Submitted queries and AI-generated responses are shown in a conversational format, with clear attribution to user and assistant roles for easy reading. The interface also features options to clear chat history, ensuring users can maintain an organized workflow throughout their literature exploration sessions.

BioTrace supports PDF, DOCX, and TXT formats. Uploaded documents are preprocessed to extract text and split into overlapping chunks of 1,000 characters with a 200-character overlap to

preserve context across chunk boundaries. These chunks are embedded using HuggingFace sentence-transformers/all-MiniLM-L6-v2 embeddings and stored in a FAISS vectorstore, which allows efficient semantic search across thousands of abstracts (Fig. 2). The vectorstore can be persisted locally or uploaded to the Hugging Face Hub, enabling collaborative sharing and version control of the knowledge base. This approach demonstrated that large-scale document embedding and retrieval can be implemented efficiently while maintaining contextual integrity. We have created a vectorstore from more than 4000 abstracts (version as of 01.11.2025), manually curated from Scopus, about eDNA focused on ecology and biomonitoring. Furthermore, with the upload function, users can add future abstracts or full-text documents to update the knowledge base.

Initial retrieval using the FAISS vectorstore identifies semantically relevant chunks for a given

query. To enhance precision, candidate documents are reranked using a CrossEncoder (cross-encoder/ms-marco-MiniLM-L-6-v2), which scores query-document pairs and prioritizes the most contextually relevant abstracts. This two-step retrieval approach reduced noise from superficially

similar but semantically irrelevant abstracts, improving the fidelity of downstream reasoning. The system's ability to combine dense retrieval with reranking highlights the importance of integrating both relevance and semantic understanding in literature exploration.

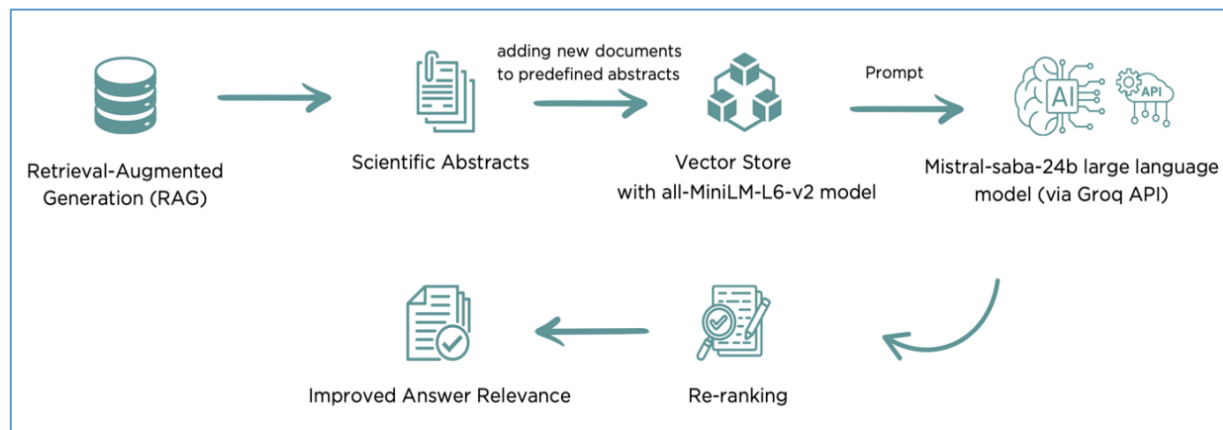


Fig 2. Workflow of the BioTrace tool.

The reranked documents serve as context for the Groq model of choice (LLama-3.1-8b-instant), which generates concise, evidence-based responses to user queries. A conversational memory buffer maintains multi-turn context, enabling iterative exploration of the knowledge base. Responses are generated with explicit citations to the original abstracts, ensuring traceability and supporting scholarly workflows. Testing on a dataset of 4,000 eDNA-related abstracts showed that users were able to retrieve relevant information across diverse topics, including species detection, environmental monitoring, and molecular techniques. The system successfully synthesized information from multiple abstracts into coherent, context-aware answers, demonstrating that conversational AI can complement traditional literature review methods.

While the system achieved strong performance in semantic retrieval and response generation, some limitations were observed. The quality of responses depends on the LLM's domain-specific understanding, occasionally producing generalized or partial answers. Processing a very large volume of data requires significant computational resources during embedding and reranking, which may limit scalability. Despite these challenges, the results suggest that integrating RAG workflows with reranking and conversa-

tional LLMs offers a practical and efficient approach to knowledge discovery in biology. Future work will focus on enhancing domain-specific LLM capabilities, optimizing memory usage for large-scale datasets, and implementing quantitative metrics to evaluate retrieval and summarization accuracy.

Conclusions

The provided tool offers some advantages over public LLM platforms such as ChatGPT, Perplexity, etc. While paid plans on these services may include advanced document analysis, they often impose strict limitations on document size and storage duration. Furthermore, the model is instructed to generate answers only from the curated database of abstracts, which will limit the hallucinations and improve the quality of the answers, compared to the general knowledge of the public LLMs. This feature makes the tool a more robust and sustainable option for academic organizations seeking greater control over their data and analytical workflow.

Acknowledgements

This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project BG-RRP-2.004-0001-C01.

References

- Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P., & Scott, A. M. (2020). A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of Clinical Epidemiology*, 121, 81-90. doi: [10.1016/j.jclinepi.2020.01.008](https://doi.org/10.1016/j.jclinepi.2020.01.008)
- Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 18(6), 1381-1391. doi: [10.1111/1755-0998.12926](https://doi.org/10.1111/1755-0998.12926)
- Cordier, T., Lanzén, A., Apothéoz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019). Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. *Trends in Microbiology*, 27(5), 387-397. doi: [10.1016/j.tim.2018.10.012](https://doi.org/10.1016/j.tim.2018.10.012)
- Digutsch, J., & Kosinski, M. (2023). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Scientific Reports*, 13(1), 5035. doi: [10.1038/s41598-023-32248-6](https://doi.org/10.1038/s41598-023-32248-6)
- Guerreiro, N.M., Alves, D.M., Waldendorf, J., Had-dow, B., Birch, A., Colombo, P., & Martins, A. F.T. (2023). Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11, 1500-1517. doi: [10.1162/tacl_a_00615](https://doi.org/10.1162/tacl_a_00615)
- Hastings, J. (2024). Preventing harm from non-conscious bias in medical generative AI. *The Lancet Digital Health*, 6(1), e2-e3. doi: [10.1016/S2589-7500\(23\)00246-7](https://doi.org/10.1016/S2589-7500(23)00246-7)
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Article 793, 9459 - 9474.
- McNichols, H., Zhang, M., & Lan, A. (2023). Algebra Error Classification with Large Language Models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13916 LNAI. doi: [10.1007/978-3-031-36272-9_30](https://doi.org/10.1007/978-3-031-36272-9_30)
- Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéoz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M. J., Filipe, A. F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Iwan Jones, J., Sagova-Mareckova, M., Moritz, C., Barquín, J., Piggott, J.J., Pinna, M., Rimet, F., Rinkevich, B., Sousa-Santos, C., Specchia, V., Trobajo, R., Vasselon, V., Vitecek, S., Zimmerman, J., Weigand, A., Leese, F., & Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, 637-638, 1295-1310. doi: [10.1016/j.scitotenv.2018.05.002](https://doi.org/10.1016/j.scitotenv.2018.05.002)
- Pham, B., Jovanovic, J., Bagheri, E., Antony, J., Ashoor, H., Nguyen, T.T., Rios, P., Robson, R., Thomas, S.M., Watt, J., Straus, S.E., & Tricco, A.C. (2021). Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow. *Systematic Reviews*, 10(1), 156. doi: [10.1186/s13643-021-01700-x](https://doi.org/10.1186/s13643-021-01700-x)
- Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *CHI EA '21: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 314, 1-7. doi: [10.1145/3411763.3451760](https://doi.org/10.1145/3411763.3451760)
- Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N., & Zhou, D. (2023). Large Language Models Can Be Easily Distracted by Irrelevant Context. *Proceedings of the 40th International Conference on Machine Learning PMLR*, 202, 31210-31227. doi: [10.48550/arXiv.2302.00093](https://doi.org/10.48550/arXiv.2302.00093)
- Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford Academics, Oxford University Press. doi: [10.1093/oso/9780198767220.001.0001](https://doi.org/10.1093/oso/9780198767220.001.0001)
- Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., & Ting, D.S.W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930-1940. doi: <https://doi.org/10.1038/s41591-023-02448-8>
- Trajanoska, M., Stojanov, R., & Trajanov, D. (2023). Enhancing Knowledge Graph Construction Using Large Language Models. *arXiv*, 2305.04676. doi: [10.48550/arXiv.2305.04676](https://doi.org/10.48550/arXiv.2305.04676)

- van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdemans, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Har-kema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D.L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2), 125-133. doi: [10.1038/s42256-020-00287-7](https://doi.org/10.1038/s42256-020-00287-7)
- Wadhwa, S., Amir, S., & Wallace, B.C. (2023). Revisiting Relation Extraction in the era of Large Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 15566-15589. doi: [10.18653/v1/2023.acl-long.868](https://doi.org/10.18653/v1/2023.acl-long.868)
- Zack, T., Lehman, E., Suzgun, M., Rodriguez, J.A., Celi, L.A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D.W., Abdulnour, R.E.E., Butte, A.J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1), e12-e22. doi: [10.1016/S2589-7500\(23\)00225-X](https://doi.org/10.1016/S2589-7500(23)00225-X)

Received: 14.10.2025

Accepted: 05.12.2025