WildInsight: a Chatbot for Wildlife Conservation Research

Anna Sokol (✉), Xianliang Zhang, and Nitesh V. Chawla

University of Notre Dame, Indiana 46556, USA
{asokol, xzhang33, nchawla}@nd.edu

Abstract. The exponential growth of machine learning (ML) and artificial intelligence (AI) presents significant benefits for wildlife conservation. However, researchers are still struggling to navigate the vast and cross-disciplinary body of literature. We introduce **WildInsight**, an LLM-powered chatbot that uses retrieval-augmented generation (RAG) to surface relevant ML applications in wildlife management. Drawing on thousands of peer-reviewed studies, WildInsight returns method overviews, species details, and geographic context answers grounded in cited sources. By bridging computational techniques and ecological practice, WildInsight accelerates evidence-based conservation decisions. Live chat is available at: http://wildinsight.lucyapps.net:1337/.

Keywords: Retrieval-Augmented Generation · Wildlife Conservation · Large Language Models · Scientific Discovery

1 Introduction

The rapid growth and complexity of ecological data presents significant challenges for traditional analytical methods. Historically, wildlife researchers relied heavily on classical statistical approaches such as regression, clustering, classification, etc. Although effective for smaller, structured datasets, these methods are increasingly inadequate given today's expansive and heterogeneous data sources, including high-volume camera-trap images, continuous acoustic recordings, satellite imagery, and genomic sequencing. Their high-dimensional, often nonlinear nature makes traditional statistics insufficient. [5].

Advancements in ML, particularly deep neural networks and LLM,offer powerful solutions, significantly enhancing the accuracy and efficiency of data-driven ecological research. State-of-the-art ML techniques, including convolutional neural networks and transformer models, now enable sophisticated tasks such as species recognition from visual and acoustic data, habitat modeling, and rapid synthesis of ecological knowledge. However, the proliferation of these ML approaches has triggered another critical issue: an exponential increase in scientific publications, making it nearly impossible for researchers to remain current. The Web of Science database, for instance, reports over one million wildlife conservation papers published in the last decade.

In response to these dual challenges of complex data analysis and literature overload, we introduce **WildInsight**, a RAG system specifically designed for ecological and wildlife research. WildInsight is built upon a rigorously curated corpus comprising the most cited peer-reviewed articles that apply ML and AI techniques in wildlife conservation research. The system's knowledge base is constructed by embedding concatenated title, abstract, and keyword texts from each publicationand storing data and metadata in a structured format. In addition, we employ query rewriting to ensure high-quality retrieval.

WildInsight's RAG framework retrieves specific, evidence-backed passages relevant to a user's query, from which an LLM generates accurate, contextually grounded responses. This methodology significantly reduces model inaccuracies and hallucinations common in generative systems.

To better understand the limitations of general-purpose research assistants for this domain, we examined prominent tools such as OpenAI's Deep Research and Google's Gemini Advanced. We observed that only a few sources suggested by these systems were peer-reviewed papers (most of the sources are links from the Internet), indicating the need for a domain-specific approach like WildInsight. Continuous input and iterative testing from domain experts, biologists, and conservation practitioners were central to refining WildInsight, ensuring that it meets real-world research and decision-making needs.

WildInsight provides citation-supported, verifiable responses through persistent DOI links, offering scientists reliable, evidence-based insights delivered at conversational speed. By integrating advanced retrieval mechanisms and robust language models, WildInsight provides a scalable and replicable framework applicable to other complex and data-rich scientific domains. Its primary users are wildlife researchers and conservation practitioners, providing rapid, evidence-based insights essential for informed ecological decision-making.

2 Related Work

RAG combines dense retrieval with sequence-to-sequence generation, enabling language models to access external knowledge sources. Lewis et al. introduced this framework, demonstrating its effectiveness in knowledge-intensive NLP tasks [3]. The system's knowledge base is constructed by embedding concatenated title, abstract, and keyword texts from each publication, and storing data and metadata in a structured format. Additionally, we employ query rewriting to ensure high-quality retrieval. Addressing the issue of hallucinations in dialogue systems, Shuster et al. showed that incorporating retrieval mechanisms significantly reduces factual inaccuracies, underscoring the importance of grounding responses in external evidence [4]. In the realm of scientific literature, Lála et al. developed PaperQA, a RAG-based agent designed to answer questions over scientific texts, highlighting the potential of RAG in facilitating scientific research [2]. While WildlifeLookup offers a chatbot for wildlife management, it lacks peer-reviewed evaluation and a comprehensive corpus. In contrast, our system, WildInsight,

provides a rigorously evaluated, citation-grounded tool tailored for wildlife conservation research [6].

3 Data Collection

We constructed WildInsight's knowledge base through a disciplined, reproducible pipeline. We queried the Web of Science Core Collection starting from a seed query—`"species" OR "wildlife"`, we expanded it with families of terms drawn from five methodological themes: (i) remote imaging (camera-traps, drones, satellites), (ii) computer vision and pattern recognition, (iii) bio-acoustics and environmental sensing, (iv) telemetry and tracking, and (v) statistical or ML models. We restricted to English-language journal and conference papers published between 1990 and 2024. Citation count ranked results and the 100,000 records were retained to prioritise influential, peer-reviewed work at the intersection of AI/ML and wildlife science. For each record we captured standard metadata (title, abstract, authors, year, DOI, keywords).

3.1 Knowledge Base Development

Our RAG approach tailored for academic paper analysis. During indexing, we merge each paper's title, abstract, and keywords into a single text chunk and create embeddings using OpenAI's text-embedding-3-large model. We selected this embedding model for its strong performance on semantic similarity tasks involving scientific texts. These embeddings are stored in Milvus, an optimized vector database. Additional metadata such as authors, publication year, and journal name are saved in separate database columns for efficient retrieval.

For the chatbot, we employ GPT-4o as the core model. Upon receiving a query, the model assesses if paper retrieval is necessary. If so, it reformulates the query to match database content (e.g., "*How is AI used to study whales?*" becomes "*Applications of AI to whale studies*"). The reformulated query is embedded and used to find the top 5 relevant papers via cosine similarity. This top-5 retrieval strategy was empirically determined to balance accuracy and concise responses during preliminary testing. The chatbot then summarizes key insights and includes DOI links from these papers in its response. If retrieval isn't needed, GPT-4o directly generates an answer without database interaction.

4 Interface and Evaluation

WildInsight features an intuitive chat interface with structured answers and *mandatory inline citations*. Typical queries-such as "*What ML methods work best for identifying birds in camera-trap images?*" "*What threats do Red Wolves face, especially related to sound and coyote hybridisation?*," or "*What are some recent ML studies on elk migration in Yellowstone?*" receive concise responses that each link to at least one DOI for seamless follow-up reading. This design

promotes efficient navigation and deeper research exploration. User interactions remain anonymous, ensuring ethical and responsible use of the system while maintaining research integrity.

To enhance usability and facilitate intuitive interaction, we have incorporated a user-friendly interface. Figure 1 illustrates the key components: a chat history panel to access past interactions, an input area to pose queries, and an output area to display WildInsight responses. This design prioritizes clarity and efficiency. Information is presented in a structured format using bullet points and headings, with source citations and DOIs readily accessible for seamless navigation to relevant publications. The interactive nature of WildInsight allows researchers to engage in a dynamic exchange, posing follow-up questions or refining queries to delve deeper into specific topics.

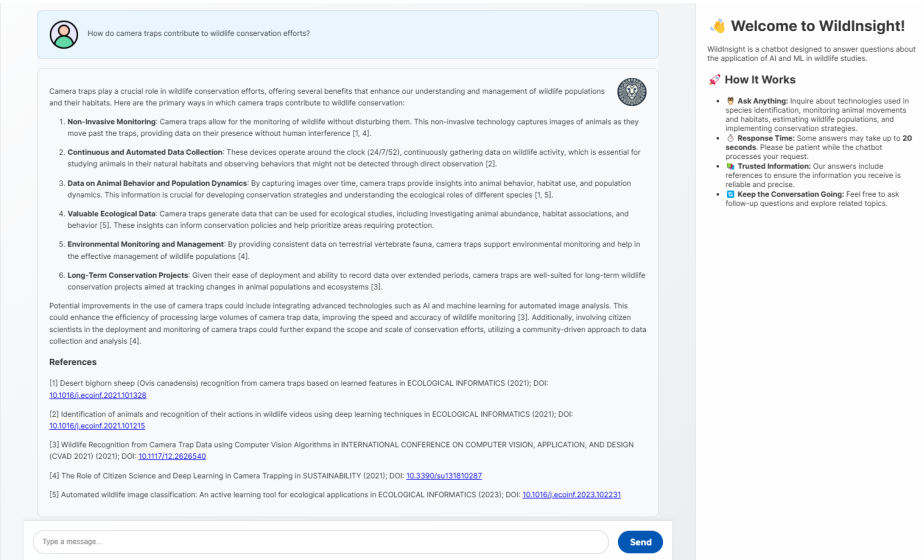

Fig. 1. Example of ChatBot Interface.

We ran 50 expert-written queries three times each; the returned paper lists overlapped with a mean Jaccard score of 0.86, showing stable retrieval on real-world questions. With no formal benchmark for this domain, the result offers a clear quantitative check on WildInsight's reliability.

5 Limitations and Future Work

WildInsight currently relies on titles, abstracts, and keywords, limiting access to full-text details. As a result, some discrepancies may occur when comparing outputs to full articles. Our evaluation reflects abstract-level consistency,

with full-text integration and analysis planned in future work. Additionally, the corpus is restricted to English-language publications, potentially excluding valuable regional studies published in other languages. We plan to expand the system's knowledge integration capabilities through: (i) implementation of knowledge graph structures representing taxonomic and ecological relationships, (ii) incorporation of grey literature including technical reports and conservation assessments, and (iii) integration with biodiversity databases such as GBIF [1] and national species inventories. Also, future evaluation frameworks will incorporate comprehensive user studies with practitioners across multiple sub-disciplines, measuring both quantitative performance metrics and qualitative assessments.

6 Conclusion

WildInsight helps researchers manage information overload in wildlife conservation by quickly finding relevant scientific literature. Our specialized RAG system connects computational methods with ecological applications through citation-based responses. This approach can be adapted to other scientific fields where researchers struggle to keep up with rapidly growing literature. We believe WildInsight provides a robust foundation for developing scalable, domain-specific research assistants that bridge computational techniques and applied ecological practices.

References

1. GBIF.Org User: Occurrence download (2025). https://doi.org/10.15468/DL.MVSESC, https://www.gbif.org/occurrence/download/0003372-250214102907787
2. Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S.G., White, A.D.: Paperqa: Retrieval-augmented generative agent for scientific research. arXiv preprint arXiv:2312.07559 (2023)
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
4. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567 (2021)
5. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., Van Langevelde, F., Burghardt, T.: Perspectives in machine learning for wildlife conservation. Nature communications **13**(1), 1–15 (2022), https://www.nature.com/articles/s41467-022-27980-y/1000, publisher: Nature Publishing Group
6. Wang, X., Yang, T., Rohr, J., Scheffers, B., Chawla, N., Zhang, X.: Wildlifelookup: A chatbot facilitating wildlife management with accessible data and insights. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining. pp. 1064–1067 (2025)