RAG Based Navigation of the World Ocean Assessment II

Zainuddin Saiyed[1], Colin Orian[1], Poppy Riddle[1], Gowri Kanagaraj[1], Geoff Krause[1], Rémi Toupin[1], Philippe Mongeon[1], Stephen Brooks[1]

[1]Dalhousie University,

Correspondence: sbrooks@cs.dal.ca

Abstract

The Second World Ocean Assessment (WOA-II) was a comprehensive evaluation of the world's oceans and climate impacts. We present a Retrieval-Augmented Generation (RAG) to query the discipline specific bibliography of the WOA-II using bibliometric metadata as the external knowledge source. Our approach uses a basic RAG pipeline with a single dense embedding, TF-IDF clustering, a single retriever with Euclidean distance similarity, and two open-source models to provide two response modes from a single non-parametric source using bibliometric abstracts as the external knowledge source. The paper details the implementation process, our evaluation of prompt development, response quality and preference, clustering, response time, and discusses limitations and future research directions. We evaluate our system by comparing the results from a variety of modes and by comparing our system's outputs to a benchmark's outputs.

1 Introduction

The current landscape of ocean research is characterized by significant challenges, including information overload due to the sheer volume of literature. Our exploratory work aims to enhance the accessibility and understanding of ocean science research by using extracted references from The Second World Ocean Assessment (WOA-II) (United Nations, 2021) and integrating the bibliometric metadata into a retrieval augmented generation (RAG) application to provide question-and-answering (Q&A) and summarization. The prototype system supports two modes of chat and summarization through a web application, enabling users to explore and derive insights from the WOA-II references corpus.

Toupin et al., (2024) previously expanded on the collection of references from the WOA-II report through the collection of cited and citing works from both Crossref and OpenAlex. The OpenAlex database (Priem et al., 2022), a comprehensive repository of scientific publications, exemplifies the information volume challenge as it contains aggregated metadata for millions of papers related to marine ecosystems.

In the remaining sections, we first review existing literature, and methodologies related to metadata and RAG applications. Section 3 details the design and implementation process, including data collection and pre-processing, clustering, system architecture, and platforms utilized. Section 4 discusses the system's interaction modes (Chat and Summarize). Section 5 presents key takeaways of the project by comparing modes and comparing our results to the OceanBench (Bi et al., 2024) benchmark. We then discusses limitations and section 6 provides recommendations for future work.

2 Prior Work

This section discusses the United Nation's World Ocean Assessments, prior work investigating RAG applications using academic text, reports, and metadata. We conclude with a few of the documented challenges for RAG that we think are pertinent to this research area.

2.1 World Ocean Assessments

The WOA-II is the major output of the second cycle of the Regular Process for Global Reporting and Assessment on the state of the marine environment, building on the work of the first World Ocean Assessment.

This report involves contributions from interdisciplinary writing teams made up of more than 300 experts, drawn from a pool of over 7,820 experts worldwide. The WOA-II aims to support decisions and actions to achieve Sustainable Development Goals, particularly Goal 14, as well as the implementation of the United Nations Decade of Ocean

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 282–290, 2025

Science for Sustainable Development, which focuses on life below water and human impacts. The topics covered in the assessment include drivers of change, current marine state and trends, pressures and impacts on the ocean, and environmental, social, and economic management (CSIRO, 2023).

2.2 Retrieval-Augmented Generation

One way to the address shortcomings with parametric knowledge of an LLM is to retrieve sources from an authoritative knowledge base outside of its training data before generating a response (Lewis et al., 2020). RAG extends a model's capabilities to specific domains or to more recent knowledge in lieu of other fine-tuning or instruction tuning methods.

Discipline specific reports such as WOA-II have similarly been used in past RAG research. Vaghefi et al., (2023) used the IPCC report as the basis for a curated selection of PDFs in their ChatClimate RAG-based chat interface in which they heuristically evaluated responses for accuracy in three conditions. Hsu et al. (2024) investigated the use of RAGs grounded in credible texts on net-zero compliance to mitigate increasing dependence on foundation models like ChatGPT by policy decision makers and as a counter to 'false and disingenuous net-zero claims'. They introduced ChatNetZero and evaluated responses by comparing responses to original texts for accuracy and asking policy experts to compare against ungrounded chatbots.

Academic text has been used with LLMs such as OpenAI's GPT models in academic search engines (Van Noorden, 2023) to meet the specific needs of students and researchers. Bi et al., (2024) created OceanGPT, a LLaMA-2 based LLM on the ocean domain, as well as OceanBench a benchmark for evaluating ocean domain responses. They used 67,633 open-access publications of parsed PDFs, exclusive of figures, tables, references, etc., as well as curated 'historically significant literature' (Bi et al., 2024). Nguyen et al. (2024) developed My Climate Advisor a RAG-based tool which returned synthesized literature responses to queries from farmers and advisors. They used scientific literature and industry grey literature on climate resilience knowledge and evaluated system performance with domain experts to assess response generation preference compared with 12 open-access and proprietary LLM models. Han, Susnjak, & Mathrani (2024) outlined a framework for RAG to automate systematic literature reviews, though they also acknowledged the challenge for retrieval.

2.3 Bibliometric metadata and abstracts

However, bibliometric metadata, either created from scientific documents or harvested from bibliometric databases, has not been used extensively. Abstracts have been used in pre-training ClimateBERT (Webersinke et al., 2022) which used research abstracts in lieu of full texts from the Web of Science, but they parsed these from full text. They have also been used in RAG pipelines with generated metadata as a retrieval filter (Poliakov and Shvai, 2024), for software comprehension (Shaik et al., 2024), or abstracts used to derive keywords for retrieval from biomedical literature used to construct systematic reviews (Li et al., 2024). Others such as Agarwal et al. (2024) created LitLLM which similarly derived keywords from seed documents and then applied multiple retrieval strategies.

2.4 Challenges for RAG

Acknowledging known challenges for RAG such as low-level perturbations (Cho et al., 2024), zero-shot precision for specialized domains (Sawarkar et al., 2025), context size (Tian et al., 2025), or struggling with real world conflicting knowledge (Marjanović et al., 2024), we wanted to know if abstracts could provide enough information to answer simple queries from users exploring the corpus. It is unknown if typos or irrelevant information in abstracts may affect responses or retrieval, or how many abstracts may be needed to provide an adequate response. It is further unknown if the abstract length compounds problems of context size, or how conflicting knowledge between two abstracts may affect an LLM's response.

3 Approach

In our prototype, RAG may be used to enhance the user's understanding of the references extracted from the WOA-II document or explore ocean related topics. Our approach uses a basic RAG pipeline with a single dense embedding, clustering, a single retriever, and two generation models to provide two response modes from a single non-parametric source using bibliometric metadata. The RAG system is implemented as per Figure 1 and will be discussed in the following sections.

3.1 Reference Data processing

DOIs from the cited works in WOA-II are used to retrieve metadata elements, such as title and ab-

stract from OpenAlex, an open-source bibliometric database. Abstracts from OpenAlex are inverted so each was reordered back to readable text. Each document is constructed from the OpenAlex identifier, title, authors, abstract, publication source, PDF URL, cited by count, publication year, publication type, and concepts. The concepts are lists of algorithmically determined labels based on title, abstract, and publication and provided by OpenAlex as an additional metadata element. Each document is saved as a plain text file for retrieval and generation and as a PDF file for fast viewing by the user resulting in a corpus of 4,723 documents.

TF-IDF is used on each text file and K-means clustering applied with the number of clusters determined by the elbow method, which involves plotting the within-cluster sum of squares against the number of clusters and selecting the point where the rate of decrease notably declines. We assign 17 clusters for our corpus and manually checked them for topical agreement.

3.2 Embedding

The user's input query is first run through the embedding model, nomic-embed-text-137M (Nussbaum et al., 2024), producing an embedding representation of that text and stored in the database. Documents with a unique chunk ID, are loaded by cluster and embedded using the same model for vector space alignment, accurate similarity measurement, and following best practices.

3.3 Database and retrieval

ChromaDB (Huber and Troynikov) was selected as it is appropriate for applications involving similarity search and retrieval of embeddings. The documents are grouped into folders, chunked as a single ID, (for convenience using the Chroma DB tools) and stored in the database. We use the built-in retriever, the Hierarchical Navigable Small World (HNSW) index and approximate nearest neighbor (ANN) search with the default L2 Euclidian distance similarity function. After some experimentation exploring minimum values, the retriever is set to k=5 for returned documents. After this, the context is assembled according to the prompt instructions.

3.4 Model selection

Our system uses Llama3 [7B] and Phi3-medium [14B] for generating the responses and the summaries along with the prompt templates for each interaction mode. The LLaMA (Touvron et al., 2023) models are open-source, instruct tuned variants and were selected for control over data, our use of copyrighted scientific literature, and reliability (Nguyen et al., 2024). We selected Phi3 (Microsoft) for similar qualities but wanted to explore if a smaller model produced acceptable responses given the brevity of abstracts. We use Ollama (Ollama, 2025) as our server framework to handle the local versions of both models and embedding model and to provide extensibility of different model usage in future. We do not quantize the models at this stage in our prototype. The query and the retrieved documents are routed through Ollama for a response from the selected model.

3.5 Interface

We use Streamlit (Streamlit), an open-source framework in Python, which was easy to use by the development team as a prototype interface for exploration. The interface includes mode selection and prescripted on and off-topic questions. In the center is the text entry for query and below were three dropdowns, one for each model response and one that includes PDFs of each retrieved source document.

3.6 FastAPI

FastAPI (Ramírez, 2025) was selected as a web framework as it supports asynchronous functions and enables functionality of our app using a local deployment across our institutional network. The API handles user input, prompt template, and interacts with Ollama for embeddings and generation.

4 System Interaction Modes

The system supports two modes of interaction: Chat mode and Summary mode, each serving distinct purposes. Chat mode is the default mode, which allows users to engage in interactive conversations and receive structured responses based on the abstracts stored in the database. Summary mode generates concise overviews of selected abstracts given a user's query.

Both modes have access to pre-scripted queries. One set is relevant to the corpus, such as "describe the current state of global fish stocks", and the other set includes incorrect queries, such as 'how do I train my dog to do tricks?". The 'irrelevant' questions were specifically included to not only demonstrate to new users how the system works, but also

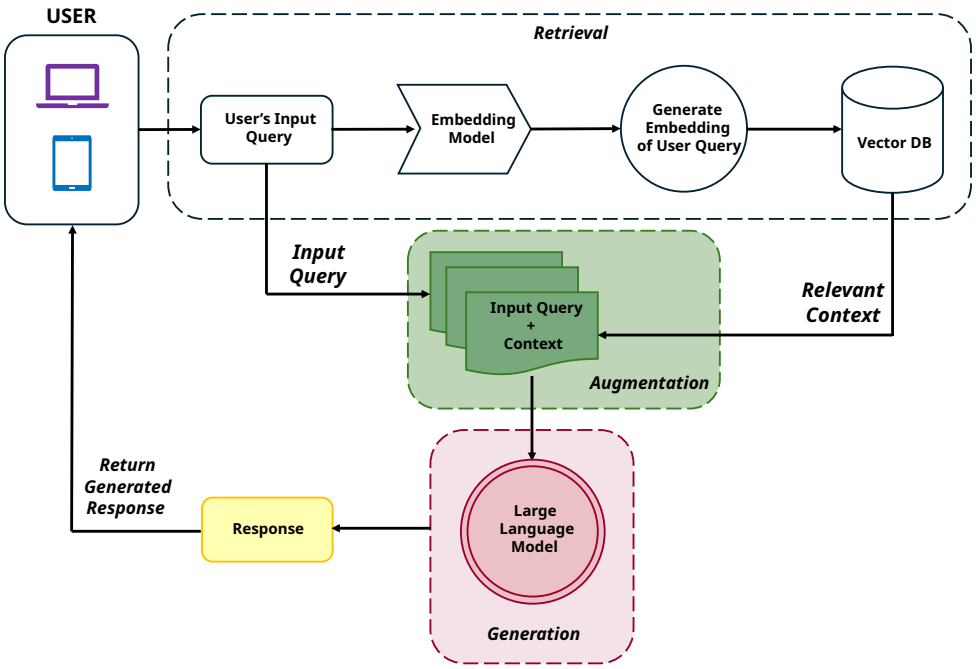

Figure 1: RAG pipeline overview.

to verify during the development that answers were only coming from the retrieved documents, and not relying upon parametric memory.

4.1 Chat Mode

Chat mode is designed to answer the user's query with concise answers given the context of the World Ocean Assessment references, from the most similar clusters. The chat prompt template (Appendix A.1) was based on the domain specific definition of the system's role, contextual relevance from the retriever, integration of the user's query and response generation instructions. The instructions consist of 8 points which emphasize strictly adhering to the facts presented in the context documents. This also helps to guard-rail the model against generating out-of-context or fictional answers to questions which are not directly relevant to the domain.

4.2 Summary Mode

Summary mode is designed to provide users with comprehensive summaries of top relevant research papers for a given user's query. This mode generates responses from multiple documents to present a cohesive overview to quickly summarize the theme, findings and research implications. The prompt template shown in Appendix A.2, is designed so that the generated summaries are comprehensive, systematically structured, and strictly based on the information available in the provided context.

This template instructs the model to analyze the context and generate a response containing key information from all relevant documents. This is to ensure the summary is comprehensive and contains insights from multiple sources to provide a complete holistic answer to the user's query. The model is instructed to generate a summary containing specific sections like Overview, Key Findings, Methodologies, Comparative Analysis, Research Implications, Future Directions, and Key Terms and Concepts. Additionally, instructions on how to cite the relevant source documents to enforce transparency and verifiability are added.

5 Evaluation and Discussion

The quality of the generated response depends on the relevance of the retrieved documents, and this is both a benefit and a challenge with RAG. We use a heuristic approach for evaluation as our team were familiar with the literature and we compared our results to a published benchmark.

From the perspective of assessing if abstracts were viable as a source, we evaluated our prompt development, response quality and model preference, the effect of cluster selection, and measured response time. Our guidelines to focus the evaluation are to assess whether bibliometric abstracts are sufficient for providing acceptable answers to simple queries, provide acceptable summaries, and

provide verifiable references. Acceptability was agreed upon to be from the perspective of expert academic users.

We used the OceanBench (Bi et al., 2024) benchmark to analyze our system. OceanBench contains over 10,000 LLM inputs related to the ocean and contains the respective outputs. Due to the size of the benchmark, it was not possible to do an exhaustive search for all relevant questions. Instead, we randomly selected a subset of questions that the LLM would likely be able to answer and questions that would be difficult for the system to answer. We also made a subset of 4 prompts by randomly selecting from all the prompts in the dataset. If a prompt was unlikely to provide a valid answer, such as the prompt: "Extract three evaluation indicators from the guide provided on the official website of the ocean environment evaluation center.", we would discard the prompt and randomly select another one until 4 valid prompts were selected [1].

5.1 Prompt Engineering Iterations

As the LLaMA family of models do not provide system prompts, there were many iterations to refine the instructions and improve the response quality of the LLM responses. For instance, in our initial base prompt, the instruction to the model was simply to answer the question based on the context retrieved from the vector database. The motivation for this base prompt is to see how the LLM generated a response to the query given the context documents given free autonomy.

For both modes, minimal prompts did not cite the referenced documents, hallucinated false documents as references, and the summaries were of varying formats. On the positive side, out of context questions were answered acceptably. The final versions, seen in Appendix A.1 and A.2, resulted in responses from both models that were contextually relevant and acceptable to users, but occasionally, the citing references were false or incorrect. The structure of in-context and out-of-context responses was consistent.

5.2 Response Acceptability and LLM Preference

We compared Llama3 [8B] and Phi3-medium [14B] for user preference regarding satisfactory responses adhering to the instructions. In both cases, we found structure was often achieved with generated responses that included point form and multiple paragraphs.

However, in some cases Phi3-medium deviates with responses structured as a single paragraph if either the context was less relevant or if the user's query was ambiguous. Otherwise, we found the Phi3-medium model was stricter in adhering to the instructions than Llama3. The generated answer by Phi3-medium was often more concise than Llama3 which many users preferred.

Overall, response quality of both models is acceptable, but some occasional errors and issues persisted. For example, as shown below, all papers were cited together as "[1],[2],[3],[4],[5]", though there was no correspondence between phrases in the sentence and the referenced documents:

> "c. Methodologies (68 words) The papers utilize a combination of empirical studies, theoretical frameworks, and case analyses to assess marine species richness, invasive species impacts on conservation planning, ecological consequences of changing climate patterns, and the behavioral aspects influencing trophic interaction [1], [2], [3], [4], [5]."

More disturbingly, the "References" section did not include the correct documents from the database:

> "References: [1] Second World Ocean Assessment [2] Comprehensive Review of Marine Biodiversity Patterns and Threats [3] Climate Change Impact on Coral Reef Systems - A Case Study Approach [4] Behavior-Mediated Trophic Interactions in Marine Mammals: An Ecological Perspective [5] Invasive Species Impacts on Marine Spatial Planning and Conservation Strategies"

Although these hallucinations were reduced and only occurred occasionally, their persistence suggests we need another mechanism in our pipeline for confirming references.

5.3 Effect of Cluster Selection

The LLM generates responses based on documents retrieved from the most semantically similar clusters. We find if a cluster is very fine-grained then the LLM is more likely to hallucinate or generate factually incorrect answers or may not have sufficient information to generate answers. This was a

[1]The subset of prompts and our system's response to the prompts can be provided upon request.

concern with using abstracts and a small corpus for the information tasks we provided. Conversely, if a cluster topic is overly broad, then the LLM generated answers which are very generic or generalized to the user's question or the domain itself.

The LLM's responses seem heavily influenced by the clustering and and may be limitations of using TF-IDF for assigning topics to works containing more than one topic. More work is needed on providing transparency of topics assigned to clusters and on exploring clustering algorithms.

5.4 Response Time

Primary factors affecting the time taken by local LLMs to generate responses are the model size, length of the generated answer, and hardware limitations. For the Llama3 models it took an average of 22 seconds for summary and 14 seconds for chat, while for the Phi3 model it took 36 seconds for summary and 16 seconds for chat, with some variation depending on the complexity of the query.

One of our exploration goals was to evaluate if we could run models locally yet not require expensive multi-GPU hardware. The hardware workstation specifications included an Intel i5-11400F [6 cores @2.6GHz] with a single NVIDIA GeForce RTX 3060 [12GB RAM], 16GB system ram [3200 MHz] and 1TB SSD storage. Both LLMs deployed used the GPU, but as we did not quantize the models, more efficiency may be gained with quantizing methods.

5.5 Benchmark Comparison

Many of our system's responses have multiple paragraphs with references and consistently use headers and lists to make the responses more human readable. The Benchmark response, however, only provide answers with a few sentences and lack references.

For general questions, such as: "Analyze the impact of marine environmental dynamics processes on the wave environment.", our system's answer has similar themes to the benchmark. For the prompt, "Categorize marine animals into several types and provide examples for each." our system's chat mode response is very similar to the benchmark response and even provides references related to protecting animals in the specific categories.

Our system does hallucinate at times and sometimes hallucinates references. Since our system provides references, a user can verify the correctness of the answer and be wary of answers that have false or no references. The benchmark outputs does not provide any references, meaning the user must trust the LLM with limited justification.

6 Conclusions and Future Work

We presented a prototype RAG application using open-source bibliometric metadata as an external knowledge source and open-source models run from a local workstation within our institutional network. Overall, we are encouraged by the results on using abstracts especially considering how few documents were retrieved (k=5) for each query.

Both models perform well with our prompts and users found responses acceptable regarding structure and response content, though there were persistent hallucinated or omitted references. When compared to the benchmark, our system's response is similar to the the benchmark with the added benefit of more details and references to justify the LLM's responses.

Future work will explore adding a verifying loop to review the references included in the response, and to continue to explore instructions that reduce false references. The effect of clusters was not evident by the user, and more work is needed to explore transparency and bibliometric clustering methods for improved topic retrieval. Response time was not perceived unfavorably by users given we were running locally, but quantizing methods may help reduce time and improve the experience.

There are several areas we look forward to continuing to explore and of course welcome any collaborators that may be interested. Currently, metadata has been curated from cited works in the WOA-II report. Future work will expand this to citing documents, greatly increasing the number and topics of works available, which may address challenges with small clusters and poor responses from highly focused clusters.

Clustering methods, particularly those around bibliometric analysis should be explored, such as using bibliometric, direct citation, co-citation, and hybrid combinations along with algorithms typically used in the bibliometrics field, such as Louvain and Leiden. Sparse, dense, and hybrid retrieval methods also need to be explored with metadata due to particular issues found in scientific texts for their effect on retrieval metrics (Cho et al., 2024; Shi et al., 2023). Post-query augmentation methods of using bibliometric measures for reranking need to be explored beyond past work using cited-

by counts. We also have concerns, from related research, about the noisiness within abstracts and titles, and much work is needed to understand the scope and effects of these characteristics on retrievers and embeddings.

The basic pipeline we have presented here will continue to be a platform for experimentation on how bibliometric metadata may serve as an openly available source of authoritative, external knowledge for future RAG applications.

Acknowledgments

Funding for this work was provided by a Canada First Research Excellence Fund grant entitled "Transforming Climate Action: Addressing the Missing Ocean".

References

Shubham Agarwal, Issam H. Laradji, Laurent Charlin, and Christopher Pal. 2024. LitLLM: A Toolkit for Scientific Literature Review. *arXiv preprint*. ArXiv:2402.01788 [cs].

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2024. OceanGPT: A Large Language Model for Ocean Science Tasks. *arXiv preprint*. ArXiv:2310.02031 [cs].

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C. Park. 2024. Typos that Broke the RAG's Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations. *arXiv preprint*. ArXiv:2404.13948 [cs].

CSIRO. 2023. Second World Ocean Assessment is afloat. Publisher: CSIRO.

Binglan Han, Teo Susnjak, and Anuradha Mathrani. 2024. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Applied Sciences*, 14(19):9103. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

Angel Hsu, Mason Laney, Ji Zhang, Diego Manya, and Linda Farczadi. 2024. Evaluating ChatNet-Zero, an LLM-Chatbot to Demystify Climate Pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 82–92, Bangkok, Thailand. Association for Computational Linguistics.

Jeff Huber and Anton Troynikov. Chroma Documentation.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yiming Li, Jeff Zhao, Manqi Li, Yifang Dang, Evan Yu, Jianfu Li, Zenan Sun, Usama Hussein, Jianguo Wen, Ahmed M Abdelhameed, Junhua Mai, Shenduo Li, Yue Yu, Xinyue Hu, Daowei Yang, Jingna Feng, Zehan Li, Jianping He, Wei Tao, and 4 others. 2024. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. *Journal of the American Medical Informatics Association*, 31(9):2030–2039.

Sara Vera Marjanović, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. *arXiv preprint*. ArXiv:2407.17023 [cs].

Microsoft. Phi Cookbook. This is a Phi Family of SLMs book for getting started with Phi Models. Phi a family of open sourced AI models developed by Microsoft. Phi models are the most capable and cost-effective small language models (SLMs) available, outperforming models of the same size and next size up across a variety of language, reasoning, coding, and math benchmarks.

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. 2024. My Climate Advisor: An Application of NLP in Climate Adaptation for Agriculture. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 27–45, Bangkok, Thailand. Association for Computational Linguistics.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic Embed: Training a Reproducible Long Context Text Embedder. _eprint: 2402.01613.

Ollama. 2025. ollama. Original-date: 2023-06-26T19:39:32Z.

Mykhailo Poliakov and Nadiya Shvai. 2024. Multi-Meta-RAG: Improving RAG for Multi-Hop Queries using Database Filtering with LLM-Extracted Metadata. *arXiv preprint*. ArXiv:2406.13213 [cs].

Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint*. ArXiv:2205.01833 [cs].

Sebastián Ramírez. 2025. FastAPI. Original-date: 2018-12-08T08:21:47Z.

Kunal Sawarkar, Shivam R. Solanki, and Abhilasha Mangal. 2025. MetaGen Blended RAG: Unlocking Zero-Shot Precision for Specialized Domain Question-Answering. *arXiv preprint*. ArXiv:2505.18247 [cs].

Kareem Shaik, Dali Wang, Weijian Zheng, Qinglei Cao, Heng Fan, Peter Schwartz, and Yunhe Feng. 2024. S3LLM: Large-Scale Scientific Software Understanding with LLMs Using Source, Metadata, and Document. In *Computational Science – ICCS 2024*, pages 222–230, Cham. Springer Nature Switzerland.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31210–31227. PMLR. ISSN: 2640-3498.

Streamlit. Streamlit.

Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. 2025. Is Relevance Propagated from Retriever to Generator in RAG? In *Advances in Information Retrieval*, pages 32–48, Cham. Springer Nature Switzerland.

Rémi Toupin, Geoff Krause, Poppy Riddle, Madelaine Hare, and Philippe Mongeon. 2024. Identifying Ocean-Related Literature Using the UN Second World Ocean Assessment Report (Dataset).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

United Nations. 2021. *The Second World Ocean Assessment: World Ocean Assessment II - Volume I & II*. United Nations.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4(1):480. Publisher: Nature Publishing Group.

Richard Van Noorden. 2023. ChatGPT-like AIs are coming to major science search engines. *Nature*. Bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Machine learning, Databases, Research data.

Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. CLIMATEBERT: A Pretrained Language Model for Climate-Related Text. *SSRN Electronic Journal*. Publisher: Elsevier BV.

A Appendix

The following are the prompts given to the respective system interaction modes. The *{context}* portion of the prompts is replaced with the abstract, title, and other metadata of all the works within the selected cluster. The *{question}* portion of the prompts is replaced with the user's query.

A.1 Chat Prompt

```
You are an expert marine scientist
assistant with comprehensive
knowledge of the Second World
Ocean Assessment
and its referenced research papers.

Context: {context}
User Query: {question}
Instructions:
1. Analyze the user's query carefully
to understand the specific information
needed.

2. Utilize the retrieved information to
identify the most relevant research
documents.

3. Synthesize information from the
relevant abstracts to formulate a
comprehensive answer.

4. Provide a clear, concise, and
factualresponse that directly
addresses the user's question.

5. Cite the relevant 'source' document
IDs in your response using the format
[Citations: [1] (the source),
[2] (the source), etc.]
and at the end list them.

6. If the question is out of context or
unclear based on the available documents,
respond with:
"That question is out of the scope of
the available documents."

7. If the information requested is not
present in the available documents,
respond with: "The information you're
seeking is not available in the
referenced documents."

8. Stick strictly to the facts presented
in the documents; do not generate
speculative answers.
```

```
Your response should be:

- Accurate and fact-based
- Directly relevant to the user's query
- Concise yet comprehensive
- Properly cited using the document
IDs and PDF citations

Answer response:
```

A.2 Summary Prompt

```
You are an expert marine scientist
assistant tasked with summarizing
research papers related to the
Second World Ocean Assessment.

Context:

{context}

Instructions:

1. Analyze the context and
synthesize their key information.

2. Create a comprehensive summary
that integrates insights from all
relevant papers in Context.

3. Structure your summary as follows:

   a. Overview (100-150 words)

  Provide a high-level summary of the main
   themes and findings across all papers.

   b. Key Findings (3-5 bullet points)

   List the most significant discoveries
   or conclusions from the papers.

   c. Methodologies (50-75 words)

   Briefly describe the primary
  research methods used across the studies.

   d. Comparative Analysis (75-100 words)

   Highlight any notable similarities
   or differences between the papers'
   approaches or results.

   e. Research Implications (50-75 words)

   Summarize the broader implications
   of these findings for marine
   science or policy.

   f. Future Directions (50-75 words)

   Identify gaps in current knowledge
   or areas for future research as
   mentioned in the papers.

   g. Key Terms and Concepts

   Define 3-5 important technical
   terms or concepts crucial to
   understanding the research.

4. Cite the relevant document Title
in your summary using the format
[1], [2], etc. and list them
at the end from the context.

5. Aim for clarity, conciseness,
and accuracy in your summary.

6. Stick strictly to the information
presented in the documents; avoid
speculation or external information.

7. If the question is out of context
or not clear based on the available
documents, respond with:
"That question is out of the scope
of the available documents."

Your summary should be:
- Comprehensive, covering the main
points from all relevant papers
- Accurate and fact-based
- Well-structured according
to the outline provided
- Properly cited using the document IDs

Summary:
```