RAG IN DOMAIN CONTEXTS: A COMPARATIVE STUDY ON DOMAIN-SPECIFIC DATASETS

Elvis Kimara, Gabriel Ferreira, Tasiful Islam & Tanim Ahmed
Department of Computer Science
Iowa State University
Ames, IA 50011, USA
{ekimara,gabferre,islam093,tanim}@iastate.edu

ABSTRACT

Large Language Models (LLMs) such as ChatGPT often generate responses that lack factual grounding, especially in specialized domains. To address this limitation, we introduce CoralAI, a Retrieval-Augmented Generation (RAG) system tailored for the marine science domain. We investigate whether domain-specific retrieval can improve LLM performance in generating accurate, verifiable answers. Our study compares multiple RAG techniques including Vanilla RAG, Chain-of-Thought (CoT) reasoning, and hybrid retrieval strategies on a curated coral-related dataset. We evaluate performance using both human-curated and LLM-generated queries, applying the RAGAS framework to assess context relevance, faithfulness, and answer quality. Preliminary findings highlight the advantages of structured retrieval and iterative reasoning in domain adaptation. This work underscores the potential of RAG systems in enhancing the trustworthiness of LLMs in specialized knowledge settings.

1 INTRODUCTION

Large Language Models (LLMs) like ChatGPT have demonstrated remarkable fluency and reasoning capabilities in general-purpose dialogue. However, they often fall short in domain-specific applications due to hallucinations, lack of verifiable sources, and limited access to up-to-date or contextually relevant information. This limitation is particularly concerning in scientific domains, where factual accuracy and citation are essential.

Retrieval-Augmented Generation (RAG) has emerged as a promising solution, enabling LLMs to ground their outputs in external, domain-specific knowledge bases. While RAG systems show potential, questions remain about their effectiveness across different retrieval strategies and query types, particularly in specialized fields.

In this study, we used CoralAI CoralX-Developers (2024), a RAG-based pipeline designed to answer queries related to coral science using a curated knowledge base. We investigate the following research questions: (1) How much does RAG improve LLM performance in a domain-specific setting compared to a base model like ChatGPT? (2) Are there measurable differences in response quality when using human-generated versus LLM-generated questions? (3) How do different RAG strategies—such as Vanilla RAG and Chain-of-Thought (CoT) reasoning—compare in terms of output accuracy?

To address these questions, we construct a dataset composed of expert-authored documents and domain-relevant questions. We apply two different RAG techniques and evaluate them using the RAGAS metric framework. Our results shed light on the strengths and limitations of domain-adapted RAG workflows and contribute insights toward building trustworthy, citation-aware LLM systems.

2 RELATED WORK

Retrieval-Augmented Generation (RAG). The idea of coupling large language models (LLMs) with a non-parametric, external knowledge store was first popularised by Lewis et al. (2020), who

showed that retrieving task-specific passages at inference time improves both factual accuracy and parameter efficiency. Subsequent work has analysed RAG's robustness and failure modes, including hallucinations that persist even when relevant documents are retrieved (Ji et al., 2023). Recent surveys provide a comprehensive taxonomy of RAG architectures and evaluation criteria (Peng et al., 2024).

Hybrid and graph-enhanced retrieval. Purely vector-based search (e.g., FAISS; Johnson et al., 2019) excels at semantic similarity but can struggle with fine-grained entity relations. Hybrid schemes therefore combine dense retrieval with symbolic representations—such as knowledge graphs—to capture both semantics and structure. Graph-augmented RAG systems have demonstrated gains on multi-hop and domain-specific QA tasks (Li et al., 2024; Sarmah et al., 2024). Our CoralAI pipeline follows this line by unifying FAISS retrieval with a NetworkX entity graph.

Reasoning-oriented prompting. Chain-of-Thought (CoT) prompting elicits intermediate reasoning steps that improve compositional generalization (Wei et al., 2022). Thompson et al. (2025) extend this idea to retrieval, iteratively issuing sub-queries (Chain-of-Retrieval Augmented Generation, or CoRAG). We adopt a similar multi-step strategy for domain questions that require explanatory chains rather than single-hop facts.

Domain adaptation vs. retrieval. Fine-tuning a foundation model on specialized corpora can inject domain knowledge but also risks catastrophic forgetting or conflicting gradients (Howard & Ruder, 2018; Ren et al., 2024). RAG sidesteps these issues by leaving model weights frozen and supplying fresh evidence at inference time, making it attractive for rapidly evolving scientific fields such as marine biology.

Evaluation of domain RAG. Beyond exact-match metrics, recent frameworks like RAGAS quantify context precision/recall, faithfulness, and answer similarity in a unified score (Patil et al., 2023). We employ RAGAS to benchmark Vanilla-RAG and CoT-RAG on both human-crafted and synthetic coral-science query sets, enabling a fine-grained analysis of retrieval and generation quality.

Together, these strands of literature motivate our study's hybrid index, citation-aware post-processing, and dual Vanilla/CoT RAG baselines for the marine-science domain.

3 Approach

Our goal is to evaluate how Retrieval-Augmented Generation (RAG) can enhance factuality in LLM responses within domain-specific contexts. In particular, we aim to quantify performance improvements over a base LLM, and analyze how retrieval and prompting strategies affect output quality. To this end, we built a hybrid RAG pipeline—CoralAI—that integrates semantic and structured retrieval methods and evaluated it using both human- and LLM-generated coral science questions.

3.1 Hybrid Knowledge Base Construction

We gathered a domain-specific knowledge base from 9 coral science PDFs CoralX-Developers (2024) and 7 out-of-domain documents. Each document was first segmented into semantically coherent chunks using spaCy-based sentence boundary detection to preserve context. Chunks were then processed in two complementary ways:

- **Vector Indexing (FAISS Facebook-AI-Research (2024)):** Each chunk was embedded using OpenAI's embedding API and indexed with FAISS Facebook-AI-Research (2024) for dense retrieval based on semantic similarity.
- **Entity Graph (NetworkX NetworkX-Developers (2024)):** We extracted domain-specific entities (e.g., coral species, climate terms) and their relationships to build a directed graph capturing structural knowledge. This allows for graph-based contextual expansion and filtering.

The resulting FAISS index and NetworkX Facebook-AI-Research (2024); NetworkX-Developers (2024) graph are saved as a unified hybrid index, enabling combined semantic and structural context retrieval for downstream tasks.

3.2 Citation-Aware Document Retrieval

To ensure that all retrieved information is verifiable, the system accounts for a citation management layer:

- APA-style citations were generated for each document using structured metadata and stored in a citation mapping file.
- The raw source paths in the FAISS Facebook-AI-Research (2024) index were replaced with the corresponding APA citations.
- During retrieval, retrieved chunks are post-processed to include these formal citations in the output, improving traceability and reliability.

3.3 RAG Strategies

We implemented and compared two primary RAG strategies to address our research questions:

- **Vanilla RAG:** A straightforward pipeline where the top-k relevant chunks from the hybrid index are retrieved and passed directly into the LLM for answer generation. This serves as a strong baseline for evaluating the benefit of retrieval augmentation.
- **Chain-of-Thought (CoT) RAG CoralX-Developers (2024):** An iterative reasoning-based approach that decomposes user queries into sub-questions (e.g., for technical definitions, causes, and implications), retrieves context for each, and generates intermediate reasoning steps before synthesizing a final answer. Prompts are structured to elicit logical reasoning, source attribution, and explicit extraction of relevant facts.

3.4 Question Set Generation and Evaluation Pipeline

To evaluate the models, we curated a balanced set of questions comprising:

- **Human-generated questions CoralX-Developers (2024):** Created by coral researchers and policymakers to serve as a gold standard.
- **LLM-generated questions:** Synthesized via few-shot prompting using coral-related documents to simulate scalable QA generation.
- **General domain questions:** Non-coral questions used to test out-of-domain robustness.

Each question was processed through both RAG pipelines. The resulting answers were evaluated using RAGAS Patil et al. (2023) metrics, which assess:

- **Context Precision and Recall:** How relevant and complete the retrieved context is.
- **Faithfulness:** Whether the generated answer adheres to the retrieved context.
- **Answer Relevancy and Similarity:** Semantic quality and closeness to ground-truth answers.

This setup allows us to systematically answer our central research questions: how RAG improves domain-specific QA, how query type affects performance, and how prompting strategies impact citation fidelity and reasoning quality.

4 Evaluation

Evaluation Summary

Our comparative analysis across base LLM, Vanilla-RAG, and CoT-RAG systems reveals three core insights. First, RAG-based systems—both Vanilla and CoT—significantly improve domain-specific

performance over the base LLM (e.g., ChatGPT). While the base LLM scored 0.79 in answer relevance and 0.85 in semantic similarity, Vanilla-RAG achieved up to 0.98 and 0.95 respectively, and CoT-RAG maintained similarly high scores, highlighting RAG's critical role in grounding responses with factual context. This improvement is most notable in context-dependent metrics such as faithfulness and context precision, where RAG pipelines leverage external documents to mitigate hallucination and deliver contextually anchored answers.

Second, a measurable quality gap exists between human- and LLM-generated questions. For both RAG variants, LLM-generated questions consistently outperform human-authored ones—achieving perfect or near-perfect scores in context recall and answer relevance (e.g., 1.0 for context recall and 0.98–0.97 for relevance), compared to lower scores on human questions (e.g., 0.94 and 0.82, respectively). This can be attributed to LLM-generated questions being designed around single document chunks, ensuring high retrievability and semantic alignment, whereas human questions are often more open-ended and contextually dispersed.

Lastly, when comparing RAG strategies, CoT-RAG generally offers modest improvements over Vanilla-RAG, particularly in context precision (0.71 vs. 0.66 on human-generated queries) and faithfulness (0.86 vs. 0.84). This suggests that Chain-of-Thought reasoning, which incorporates symbolic graph traversal, enhances multi-hop reasoning and reduces retrieval noise, especially for complex or ambiguous queries. But the scores for semantic similarity and answwr relevancy were slightly below than the Vanilla rag indicating the internal iterative subquery process can shift the final answer slightly from the ground truth resulting in slightly wea.

Overall, these results confirm that RAG significantly enhances LLM performance in specialized domains, with LLM-generated QA inputs and hybrid reasoning strategies further pushing the boundaries of retrieval accuracy and response faithfulness.

Summary of Findings

- **How much does RAG improve LLM performance in a domain-specific setting?** RAG improves LLM performance significantly, boosting relevance from 0.79 to 0.82 and similarity from 0.85 to 0.90.
- **Are there measurable differences between human-generated and LLM-generated questions?** LLM-generated questions outperform human-authored ones in all metrics due to better alignment with retrievable content.
- **How do Vanilla-RAG and Chain-of-Thought (CoT) RAG compare?** CoT-RAG slightly outperforms Vanilla-RAG, especially in context precision and faithfulness, thanks to structured reasoning.

	Context Precision			**Context Recall**			**Faithfulness**			**Answer Relevance**			**Answer Similarity**		
	Human	LLM	Gen	Human	LLM	Gen	Human	LLM	Gen	Human	LLM	Gen	Human	LLM	Gen
Base LLM	-	-	-	-	-	-	-	-	-	.79	.99	.97	.85	.93	.92
Vanilla-RAG	.66	.94	1	.95	1	.92	.84	.89	.76	.82	.98	.96	.90	.95	.95
CoT-RAG	.71	.93	1	.94	.98	.93	.86	.92	.81	.81	.97	.93	.88	.94	.92

Table 1: Performance comparison across methods (Base LLM, Vanilla-RAG, CoT-RAG), question types (Human, LLM, General) and evaluation metrics.

5 Conclusion

References

CoralX-Developers. Coralx project: Domain-specific coral science dataset, expert-curated question set, and rag systems. `https://coralxfoundation.com`, 2024. Accessed: 2025-05-14.

Facebook-AI-Research. Faiss: A library for efficient similarity search. `https://github.com/facebookresearch/faiss`, 2024. Accessed: 2025-05-14.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*, 2018.

Xinyi Ji, Chenxin An, Yujia Qin, and Xiang Ren. A survey on hallucination in large language models. *arXiv*, 2311.05232, 2023.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. doi: 10.1109/TBDATA.2019.2921572. Originally released as arXiv:1702.08734.

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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020. arXiv:2005.11401.

Keqing Li, Yina Zhang, and Fei Lv. Gnn-rag: Graph neural retrieval for large language model question answering. *arXiv*, 2405.20139, 2024.

NetworkX-Developers. Networkx: Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. `https://networkx.org/`, 2024. Accessed: 2025-05-14.

Suraj Patil et al. Ragas: Retrieval augmented generation assessment. `https://github.com/explodinggradients/ragas`, 2023. Accessed: 2025-05-14.

Bo Peng, Chao Li, Rui Zhang, and Wei Zhao. Graph retrieval-augmented generation: A survey. *arXiv*, 2408.08921, 2024.

Xiang Ren, Jinhao Jiang, and Junyi Li. Pitfalls of domain adaptation for large language models. *arXiv*, 2407.10804, 2024.

Chiranjit Sarmah, Bo Peng, and Sajad Tahmasebi. Hybridrag: Integrating knowledge graphs and vector retrieval for enhanced qa. *arXiv*, 2408.04948, 2024.

Michael Thompson, Rui Zhang, and Aman Singh. Chain-of-retrieval augmented generation. *arXiv*, 2501.14342, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, 2201.11903, 2022.

A Artifact Details

A.1 Code Repository

A.2 README Checklist