DEVELOPING A COMPREHENSIVE BENCHMARK DATASET FOR RAG APPLICATIONS IN SCIENTIFIC KNOWLEDGE DISCOVERY

Mahira Ibnath Joytu
Project Assistant
Information Technology
Faculty of Science and Engineering
Åbo Akademi University

Md Raisul Kibria
Doctoral researcher
Information Technology
Faculty of Natural Sciences and Engineering
Åbo Akademi University

Sebastien Lafond
Professor
Information Technology
Faculty of Science and Engineering
Åbo Akademi University

Table of Contents

1. Introduction

1.1 Overview

Due to significant improvements in the capabilities and size of autoregressive language models, they have been implemented in a wide range of application areas for intelligent solutions with minimal human involvement. Hallucinations, a major issue that such generative large language models (LLMs) may suffer from, can be addressed through Retrieval-augmented generation (RAG), where relevant context is retrieved and integrated into the LLM's generation process. In the scientific domain, specifically in tasks such as literature review, a huge amount of information is required to be processed after finding relevant sources before coming to factual conclusions. RAG agents have already been demonstrated to significantly accelerate this process, making the agent as effective as human experts [1]. However, existing benchmarks for such agents are concentrated in certain fields (such as biomedicine) and lack comprehensive assessment criteria to evaluate retrieval and generation. In this mini-project, we aim to generate a benchmark dataset for question-answering RAG agents on scientific information, focusing on human-likeness for retrieval and four criteria – robustness, grounding, knowledge expansion, and counterfactual identification for generation, following Chen et al. [2]

1.2 Objective

The main objective of this dataset is to provide a multi-factorial test suite to rank RAG agents with different LLMs for retrieval, generation quality, and accuracy. By focusing on human-like retrieval and evaluating generation along criteria like robustness, grounding, knowledge expansion, and counterfactual identification, this dataset aims to support the development of more reliable and versatile RAG agents. Unlike existing benchmarks concentrated in a specific field, this dataset covers a broader range of scientific disciplines, ensuring a more comprehensive assessment of the agents' performance across various domains. Ultimately, it will facilitate a deeper understanding of how various LLMs perform in scientific question-answering tasks.

2. Literature Review

Several benchmarking efforts have been initiated to evaluate the efficacy of RAG systems comprehensively. The Retrieval-Augmented Generation Benchmark (RGB), proposed by Chen et al. [2], evaluates LLMs based on four core abilities: noise robustness, negative rejection, information integration, and counterfactual robustness. Their study shows that while LLMs demonstrate some noise robustness, they struggle significantly with rejecting incorrect information, synthesizing data from multiple sources, and identifying factual inconsistencies. However, the RGB benchmark primarily focuses on biomedicine and lacks evaluation across other scientific domains, limiting its generalizability.

Similarly, the Comprehensive RAG Benchmark (CRAG) introduced by Yang et al. [3] aims to bridge the gaps in existing benchmarks by incorporating a wide range of factual question-answer pairs across multiple domains, such as finance, sports, and movies. CRAG evaluates RAG systems on their ability to handle dynamic, low-popularity, and complex queries. The findings reveal that while retrieval augmentation improves accuracy from 34% to 44%, hallucination remains a major challenge, with state-of-the-art solutions achieving only 63% accuracy. Despite its comprehensiveness, CRAG lacks coverage of highly specialized scientific disciplines, focusing instead on general knowledge domains.

In the context of scientific research, the PaperQA system by Lála et al. [1] demonstrates the potential of RAG agents to match or exceed human performance in literature-based question-answering tasks. PaperQA retrieves relevant full-text scientific articles, synthesizes information, and generates answers with citations. The system outperforms other LLM-based agents on benchmarks like PubMedQA and the newly introduced LitQA dataset. Notably, PaperQA reduces hallucination rates and exhibits improved knowledge boundaries by correctly identifying when insufficient evidence is available. However, PaperQA focuses exclusively on biomedical literature, limiting its applicability to other scientific domains, and it relies heavily on the quality of the retrieval mechanism, which may fail when relevant papers are inaccessible.

These benchmarks and systems underscore the importance of robust evaluation criteria for RAG agents, especially in scientific domains where precision is crucial. Current RAG models are effective in mitigating hallucinations to some extent but require improvements in integrating multi-source data and rejecting misleading information. The development of broader benchmarks covering diverse scientific disciplines, as outlined in the objectives of this project, will be instrumental in enhancing the reliability and versatility of RAG systems.

In conclusion, while RAG provides a promising approach to improving LLM accuracy, continued development of comprehensive benchmarks and advanced retrieval techniques is essential to address existing limitations and achieve human-like performance in scientific information processing.

3. Methodology

3.1 Defining Areas, Sub-Areas and Topics

The goal of the project is to create a comprehensive benchmark dataset that will span across various domains to evaluate RAG systems. Hence, a total of 10 areas have been selected to be included in the dataset ranging from Arts and Humanities, Environment to Physics. The areas have been defined following the Nature journals. Each area has been divided into specific research sub-areas and further divided into more focused research topics such as the sub-area Electronics, within Physics, includes the topic Energy Harvesting. The defined areas with their corresponding sub-areas and topics are described in this section.

1. **Physics**

Sub-Area	Topics
1.1 General Physics:	a. Photonics
	b. Thermodynamics
1.2 Electronics:	a. Energy Harvesting

Physics is central to technological and scientific innovation. In the sub-area of General Physics, topics like Photonics and Thermodynamics explore fundamental principles with applications in energy, communication, and materials science. Photonics, for example, focuses on generating and manipulating light, which is crucial for developing optical technologies and communication systems. Electronics research, particularly in Energy Harvesting, investigates methods for capturing and utilizing ambient energy sources, contributing to advancements in renewable energy and low-power devices. These topics are rich in theoretical models, experimental data, and practical applications, making them suitable for evaluating the depth and accuracy of RAG agents.

2. Chemistry

Sub-Area	Topics
2.1 Forensic Chemistry:	a. Gunshot Residue
2.2 Contamination:	a. Microplastics

Chemistry is a practical and analytical science that underpins forensic investigations and public health. In the sub-area of Forensic Chemistry, topics like Gunshot Residue play a key role in criminal investigations, helping to determine the presence of firearms and reconstruct crime scenes. This field requires precise analytical techniques and detailed reporting. Another important area is Contamination, specifically Microplastics. The study of microplastics in food and water addresses growing concerns about environmental pollution and its effects on human health. These topics demand accuracy and attention to detail, making them ideal for assessing the ability of RAG agents to handle analytical and experimental data.

3. Computer Science

Sub-Area	Topics
3.1 Artificial Intelligence:	a. Explainable AI
	b. Neural Networks
	c. Machine Learning
3.2 Internet of Things (IoT):	a. Security of IoT Systems

Computer science is a foundational domain for modern technological advancements. In the sub-area of Artificial Intelligence, topics such as Explainable AI, Neural Networks, and Machine Learning are critical for ensuring transparency, fairness, and efficiency in AI models. Explainable AI addresses the need for interpretable decision-making processes, which is increasingly important in sectors like healthcare, finance, and law. Neural networks and machine learning are core to AI development, driving innovations in fields such as natural language processing, computer vision, and autonomous systems. Another important sub-area is the Internet of Things (IoT), with a focus on the Security of IoT Systems. As more devices become interconnected, securing these systems against cyber threats is paramount. Research in these topics provides a blend of theoretical frameworks, algorithms, and practical applications, making them ideal for assessing the retrieval and generation capabilities of RAG agents.

4. Biochemistry, Genetics, and Molecular Biology

Sub-Area	Topics
4.1 Cancer Research:	a. Leukemia
	b. Lung Cancer
4.2 Protein Structure:	a. AlphaFold

This domain is at the forefront of medical and molecular research, making it essential for evaluating the precision and depth of RAG agents. In the sub-area of Cancer Research, topics such as Leukemia and Lung Cancer involve extensive research on disease mechanisms, diagnostics, and therapeutic strategies. The volume of literature in this field is immense, reflecting the global urgency to combat cancer through innovative treatments and personalized medicine. Accurate retrieval of clinical trials, genetic studies, and treatment outcomes is essential for meaningful insights. Another critical sub-area is Protein Structure, particularly advancements like AlphaFold that predict protein folding with remarkable accuracy. This research has transformed drug discovery, enabling scientists to understand protein interactions at a molecular level. The complexity and specificity of these topics challenge RAG agents to process and synthesize detailed biochemical data effectively.

5. Agricultural and Biological Sciences

Sub-Area	Topics
5.1 Food Science:	a. Food Chemistry
5.2 Protein Structure:	a. Animal Ecology

Agricultural and biological sciences are critical for addressing global challenges in food security, sustainability, and ecosystem preservation. In the sub-area of Food Science, topics such as Food Chemistry involve studying the molecular composition, processing methods, and safety of food products. This research is essential for ensuring public health, improving nutritional quality, and addressing global food safety concerns. The extensive body of literature on additives, contaminants, and preservation techniques provides rich data sources for RAG agents to retrieve and synthesize information. Another key sub-area is Ecology, Evolution, Behavior, and Systematics, with a focus on Animal Ecology. This sub-field examines interactions within ecosystems, population dynamics, and species behavior. Research in animal ecology is crucial for conservation efforts, biodiversity management, and understanding the impacts of climate change on wildlife. The topics are both interpretative and data-driven, making them ideal for evaluating the ability of RAG agents to handle detailed ecological datasets and case studies.

6. Psychology

Sub-Area	Topics
6.1 Social Psychological and Personality Science:	a. Substance Use
	b. Mental Health
6.2 Gambling Studies:	a. Gambling Behavior
	b. Cyberpsychology

Psychology offers rich insights into human behavior, cognition, and mental health. In the sub-area of Social Psychological and Personality Science, topics like Substance Use and Mental Health address critical societal issues such as addiction, anxiety, and depression. The increasing prevalence of mental health concerns worldwide has led to a significant body of research on interventions, therapies, and public health strategies. Additionally, Gambling Studies and Cyberpsychology explore behaviors influenced by technology, risk-taking, and digital environments. Gambling behavior studies examine patterns of addiction and the psychological impacts of gambling, while cyberpsychology investigates how digital interactions affect mental and emotional well-being. These topics are socially relevant, offering opportunities to evaluate how well RAG agents retrieve and synthesize complex psychological research.

7. Health Science

Sub-Area	Topics
7.1 Health Science:	a. COVID-19
	b. Mpox

Health science is essential for addressing global health challenges and improving public well-being. In the sub-area of Public Health, topics such as COVID-19 and Mpox reflect the ongoing need for accurate and up-to-date health information. The COVID-19 pandemic has generated vast research on virology, vaccines, and public health responses, while Mpox (formerly known as monkeypox) highlights the importance of monitoring emerging infectious diseases. These topics are critical for understanding disease prevention, health policy, and medical interventions. The extensive body of literature in public health provides a dynamic and highly relevant dataset for evaluating RAG agents' retrieval and synthesis capabilities.

8. Environment

Sub-Area	Topics
8.1 Grasslands:	a. Change in Grassland Use (Europe)
8.2 Pollution:	a. Pesticide Residue
8.3 Climate Change:	a. CO_2 Emissions
	b. Wildfire Exposure

The environmental sciences address pressing global challenges related to sustainability, pollution, and climate change. In the sub-area of Grasslands, research on the Change in Grassland Use in regions like Europe examines the impacts of agriculture, urbanization, and conservation efforts on these ecosystems. Grasslands play a vital role in biodiversity and carbon sequestration, making this research crucial for environmental policy. Pollution studies, such as those on Pesticide Residue, focus on the environmental and health impacts of agricultural chemicals. This area is essential for understanding food safety, water quality, and ecosystem health. The sub-area of Climate Change covers topics like CO_2 Emissions and Wildfire Exposure, both of which are critical for mitigating environmental risks and understanding the effects of global warming. These topics provide dynamic and data-rich challenges for RAG agents to process and retrieve environmental research effectively.

9. Arts and Humanities

Sub-Area	Topics
9.1 Political Geography:	a. Arctic Geopolitics
	b. Settler Colonialism
	c. Energy Transition

The domain of arts and humanities provides a platform for understanding socio-political dynamics and human experiences. In Political Geography, topics such as

Arctic Geopolitics, Settler Colonialism, and Energy Transition are especially relevant. Arctic geopolitics explores territorial claims, resource extraction, and the geopolitical implications of melting ice caps. This field is dynamic and ever evolving, influenced by climate change and international relations. The study of settler colonialism delves into historical and ongoing processes of colonization, examining land rights, cultural impacts, and social justice. Energy transition focuses on the shift from fossil fuels to renewable energy, touching on economic, environmental, and policy challenges. These areas are rich in historical data, policy analysis, and contemporary debates, offering a diverse set of resources for benchmarking the retrieval capabilities of RAG agents.

10.Business, Management, and Accounting

Sub-Area	Topics
10.1 VR in Business:	a. Hospitality
	b. B2B Marketing
10.2 Marketing:	a. Sustainability in Business
	b. Virtual Influencers

The business domain is evolving rapidly due to technological advancements and changing market dynamics. In the sub-area of VR in Business, topics such as Hospitality and B2B Marketing illustrate how virtual reality enhances customer experiences and training programs. This field provides numerous case studies on the implementation of VR technologies in service industries and corporate environments. In Marketing, topics like Sustainability in Business and Virtual Influencers reflect contemporary trends where companies integrate environmental responsibility and digital innovation into their strategies. Sustainability-focused marketing emphasizes eco-friendly practices and corporate social responsibility, while virtual influencers represent a new wave of digital engagement. The literature in these areas is rich with real-world applications, making them ideal for evaluating the ability of RAG agents to retrieve and process business-related information.

3.2 Defining question criteria

To comprehensively evaluate the capabilities of Retrieval-Augmented Generation (RAG) agents, a structured framework for question difficulty is essential. We have developed a four-level difficulty rating that categorizes questions based on the complexity of retrieval and synthesis required, as well as the challenges posed to the model's robustness and reasoning abilities.

3.2.1 Level 1: Single-Line Synthesis or Direct Retrieval

Level 1 questions are the simplest, where answers can be found directly within a single line of text in the source paper or require minimal synthesis. These questions test the basic grounding of the RAG agent, evaluating its ability to identify and retrieve factual information accurately. For example, a question like “What year was the experiment conducted?” or “What is the accuracy score of the model?” fall into this category.

3.2.2 Level 2: Multi-Line Synthesis or Inference

The questions of this difficulty level require the agent to synthesize information from multiple lines within a single passage or make inferences based on the provided content. These questions test the agent's ability to perform knowledge expansion and robustness by integrating details spread across several sentences. For instance, a question such as “In the paper "Outcome of Adolescents and Young Adults with Acute Lymphoblastic Leukemia in a Single Center in Brazil," did the majority of patients belong to the adolescents and young adults (AYA) group and receive the Berlin-Frankfurt-Münster (BFM) protocol?” demands that the agent piece together information scattered throughout a paragraph.

3.2.3 Level 3: Multi-Passage Synthesis and Negative Examples

Level 3 questions increase in complexity by requiring synthesis from multiple passages within the paper. Additionally, this level includes negative example questions, where the correct answer is intentionally absent from the document. These questions are designed to trick the RAG agent and test its counterfactual identification and robustness. For example, a question like “How many Super Bowls did Messi win?” does not have an answer, challenging the agent to avoid generating inaccurate information.

3.2.4 Level 4: Cross-Document Synthesis

Questions belonging to this difficulty level are the most challenging, requiring the agent to synthesize information from two or more related papers. These questions test the highest levels of knowledge expansion, grounding, and robustness. They evaluate the agent's ability to integrate diverse datasets, compare findings, and generate comprehensive answers. An example of a Level 4 question might be “How do the conclusions about climate change impacts in Paper A compare with the findings in Paper B?”

3.2.5 Distribution of difficulty level

The difficulty distribution of the dataset is set to have a predominance of easier questions. Level 1 questions, which represent the lowest difficulty account for approximately 44.3% of the total dataset. Level 2 questions, representing moderately easy difficulty, make up 25.5% of the dataset. Level 3 questions, which are categorized as moderately difficult, constitute 18.9% of the total. Finally, Level 4 questions, representing the highest difficulty level account for 11.3% of the dataset. The skew toward lower difficulty level ensures that the dataset is structured to include a greater proportion of questions that are accessible and easier to answer for the RAG model, while progressively more challenging questions are less prevalent. We want to judge the model on all difficulty levels fairly, thus, even though higher difficulty questions are less frequent, our priority remains to assess the model’s ability to answer easier questions first and then make the questions progressively harder. This will allow us to assess the model’s performance across all levels and then make necessary observations and adjustments to fine-tune the model to be able to handle difficult questions more robustly. The following chart visualizes the distribution of questions of each difficulty level:

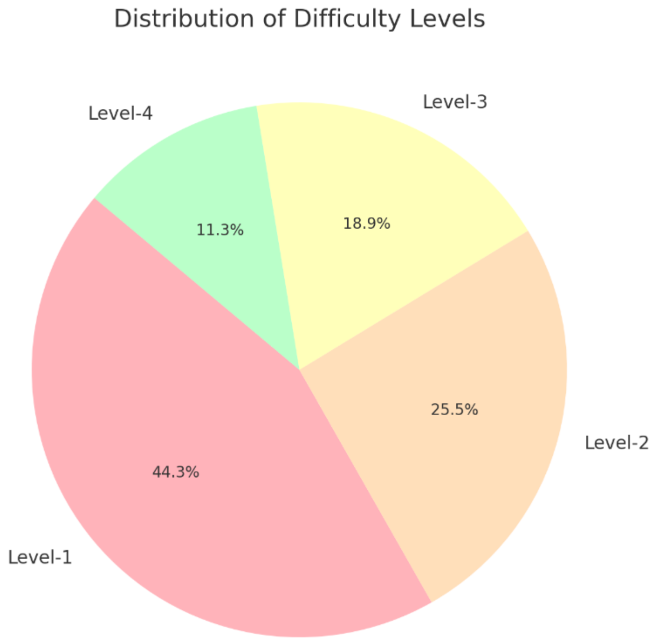

Fig- 1: Distribution of difficulty levels

3.3 TOOLS AND TECHNOLOGY

To publish and store our dataset, we are using Zenodo which is an open-access research repository developed by CERN and funded by the European Commission. It allows researchers to upload, share, and preserve research outputs such as datasets, publications, software, and presentations. It assigns a Digital Object Identifier (DOI) to each upload, ensuring long-term accessibility and citation while also supporting a wide range of file types and integrates with platforms like GitHub for software preservation.

The dataset has been populated and formulated using Microsoft's Excel software which allowed us to seamlessly add, edit and maintain the dataset with its numerous functionalities.

The majority of the papers have been sourced from ScienceDirect which is a leading platform for accessing scientific and technical research articles, operated by Elsevier. It provides a vast collection of peer-reviewed journals, books, and conference proceedings across a wide range of disciplines making it an optimal source of papers for us. ScienceDirect houses high-quality, up-to-date research, offering advanced search and filtering tools for efficient discovery. It ensures access to full-text articles and promotes reliable sourcing for academic and professional research. For articles without full-text access, we used Åbo Akademi University's VPN tunnel to unlock them. Other sources of papers include Nature, Google Scholar and Arxiv.

4. DATASET

The benchmark dataset has been curated to include 106 questions with the following columns:

i. Paper: Unique 4-dimensional index number of the source paper E.g. [1][0][1][1]

ii. Question: The question that will be the input of the RAG model

iii. Answer: The one-word or numerical answer to the respective question i.e.

iv. Reference: The source text of the answer

v. Difficulty: The difficulty rating of the question between the range of 1-4

The paper index has been included to verify if the RAG model is sourcing the answer from the desired paper or not. The inclusion of the reference column allows us to formulate relevant answer options to provide to the model.

Additionally, we have formulated another dataset to include all the source papers which has the following columns:

i. Area: The broad subject area of the paper

ii. Sub-Area: The research Sub-Area of the paper

iii. Topic: The focused research topic of the paper

iv. Paper: The name of the paper

v. Index: Unique 4-dimensional index of the paper E.g. [1][0][1][1]

vi. Link: The URL linking to the page where the paper has been sourced from

5. CONCLUSION

With this project we have successfully developed a comprehensive benchmark dataset for evaluating Retrieval-Augmented Generation (RAG) agents in scientific knowledge discovery. The dataset spans 10 broad scientific areas, each subdivided into sub-areas and specific research topics, ensuring coverage across diverse disciplines, including Physics, Chemistry, Computer Science, Business, Agriculture, and more. A total of 108 questions were curated and categorized by four difficulty levels to assess the robustness, grounding, knowledge expansion, and counterfactual identification capabilities of RAG models.

The dataset design emphasizes human-like retrieval and high-quality generation, with the goal of addressing gaps in existing benchmarks, which often focus narrowly on fields like biomedicine. By incorporating papers from reliable sources such as ScienceDirect and Nature, the dataset ensures that the retrieval tasks are grounded in credible, peer-reviewed research. The structured 4-dimensional indexing system enables precise referencing and systematic evaluation of RAG performance across multiple domains.

This benchmark provides a valuable tool for the ongoing development and fine-tuning of RAG agents, facilitating advancements in their ability to process, retrieve, and synthesize scientific information accurately. Future work can expand the dataset by incorporating additional disciplines and refining evaluation criteria to further enhance the versatility and reliability of RAG systems in scientific applications.

6. REFERENCES

[1] J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, and A. D. White, “PaperQA: Retrieval-augmented generative agent for scientific research,” arXiv preprint arXiv:2312.07559, Dec. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.07559

[2] J. Chen, H. Lin, X. Han, and L. Sun, “Benchmarking large language models in retrieval-augmented generation,” arXiv preprint arXiv:2309.01431, Sep. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.01431

[3] X. Yang et al., “CRAG – Comprehensive RAG Benchmark,” arXiv preprint arXiv:2406.04744, Jun. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.04744