# WeQA: A Benchmark for Retrieval Augmented Generation in Wind Energy Domain

**Rounak Meyur, Hung Phan, Sridevi Wagle, Jan Strube, Mahantesh Halappanavar, Sameera Horawalavithana, Anurag Acharya, Sai Munikoti**
Pacific Northwest National Laboratory
Richland, WA 99354
{rounak.meyur, hung.phan, sridevi.wagle, jan.strube, mahantesh.halappanavar, yasanka.horawalavithana, anurag.acharya, sai.munikoti}@pnnl.gov

## Abstract

Wind energy project assessments present significant challenges for decision-makers, who must navigate and synthesize hundreds of pages of environmental and scientific documentation. These documents often span different regions and project scales, covering multiple domains of expertise. This process traditionally demands immense time and specialized knowledge from decision-makers. The advent of Large Language Model (LLM)s and Retrieval Augmented Generation (RAG) approaches offer a transformative solution, enabling rapid, accurate cross-document information retrieval and synthesis. As the landscape of Natural Language Processing (NLP) and text generation continues to evolve, benchmarking becomes essential to evaluate and compare the performance of different RAG-based LLMs. In this paper, we present a comprehensive framework to generate a domain relevant RAG benchmark. Our framework is based on automatic question-answer generation with Human (domain experts)-AI (LLM) teaming. As a case study, we demonstrate the framework by introducing WeQA, a first-of-its-kind benchmark on the wind energy domain which comprises of multiple scientific documents/reports related to environmental aspects of wind energy projects. Our framework systematically evaluates RAG performance using diverse metrics and multiple question types with varying complexity level, providing a foundation for rigorous assessment of RAG-based systems in complex scientific domains and enabling researchers to identify areas for improvement in domain-specific applications.

## 1 Introduction

In recent years, the advancements in LLM have revolutionized various natural language processing tasks, including text and response generation. However, text generation using LLM often encounters challenges such as generating irrelevant or incoherent outputs, perpetuating biases ingrained in the training data, and struggling to maintain context and factual accuracy (Wu et al., 2024). These issues pose significant obstacles to achieving human-level performance in automated text generation systems. RAG effectively mitigates these common challenges by incorporating retrieved information to enhance coherence and factual accuracy, thus minimizing the generation of fictitious or irrelevant content (Gao et al., 2024; Lewis et al., 2021). Furthermore, concurrent works suggest RAG is the most sought approach for adapting models towards accelerating repetitive and data intensive tasks in niche scientific domain such as nuclear, renewable energy, environmental policy, etc. (Munikoti et al., 2024a,b; Phan et al., 2023). While RAG-based systems have demonstrated promising capabilities in streamlining document analysis tasks across various professional domains, their integration into critical decision-making processes like permitting wind energy projects remains constrained due to legitimate concerns about trust and reliability.

In this work, we create benchmarks to assess RAG-based LLM performance in the domain of permitting wind energy projects. Environmental Impact Statements (EIS) represent the cornerstone documentation within this permitting landscape, serving as comprehensive analyses that evaluate the potential environmental consequences of proposed wind energy developments. These documents play a pivotal role in promoting informed decision-making by ensuring transparency and incorporating diverse stakeholder perspectives into the approval process (Bond et al., 2024). By providing detailed evaluations of environmental effects, alternatives analysis, and mitigation measures, EIS documentation facilitates the responsible development of wind energy infrastructure while building public trust at the same time.

As RAG-based LLMs gain traction for domain-specific applications such as wind energy permitting, their effectiveness must be rigorously assessed



*Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 239–251

through robust benchmarks to ensure its practical utility and reliability (Chen et al., 2023a). Establishing high-quality benchmarks is essential to evaluate their abilities to perform regulatory-focused reasoning, accurately interpret complex EIS documents, and support logical deductions grounded in the documents. Such benchmarks facilitate systematic assessment of how well RAG-based LLMs can handle the nuanced requirements of the domain (Xiong et al., 2024). A robust evaluation framework allows researchers and practitioners to investigate the impact of retrieval strategies, model architectures, and training data, on the performance of RAG, while building confidence in automated tools for critical environmental decision making (Ray, 2023).

In benchmarking RAG for wind energy project permitting applications, it is crucial to evaluate its performance across a diverse set of questions that reflect the complexity and variability of real-world permitting scenarios (Lyu et al., 2024). A set of well curated and diverse questions enable a comprehensive assessment of RAG's ability to interpret EIS documents, analyze environmental impacts, evaluate regulatory compliance, and generate coherent responses to permitting-related queries that practitioners encounter during wind energy project review processes. To generate such questions, automated methods leveraging NLP techniques can be employed, including rule-based approaches that capture language patterns from relevant documents, template filling methods that incorporate wind energy terminologies, and neural network-based models that can efficiently create diverse question sets by leveraging the semantic relationships inherent in EIS and other documents related to wind energy projects.

Human-curated questions offer a level of linguistic richness and contextual relevance that may be challenging to achieve solely through automated generation methods, particularly in specialized domains such as wind energy project permitting (Zhang et al., 2024). By leveraging human expertise and domain knowledge, curated question sets can encompass a broader spectrum of linguistic variations, domain-specific considerations, and nuanced semantics (Ribeiro et al., 2020), providing a more comprehensive evaluation of RAG's performance across diverse scenarios and applications (Thakur et al., 2021). Combining automated generation with human curation for benchmarking RAG offers a synergistic approach to ensure both efficiency and quality in question sets. This hybrid approach leverages the strengths of both automated and human-driven processes, that provide efficient and robust evaluation metrics for RAG's performance.

In this work, we present a hybrid workflow to benchmark RAGs, which combines rapid question generation through automated methods, augmented with properly designed human prompts to generate diverse set of questions. Our proposed benchmarking framework is used to generate questions from EIS and other research documents related to environmental impact of wind energy projects. The extensive question-answer dataset serve as a tool to evaluate the performance of RAG-based LLMs, which are designed to answer queries related to these extensive and comprehensive documents. Given the vast amount of information contained in these documents, manually reviewing them is impractical, making RAG-based LLMs essential for generating accurate responses to specific queries. Our benchmarking framework assesses the effectiveness of these models in accurately retrieving and responding to queries, ensuring that they can reliably process and provide relevant information from the documents.

**Contributions** The paper introduces a novel benchmark dataset for question-answering (QA) task in a specific domain and also proposes a generic framework to evaluate the RAG-based LLM responses to different entries in the benchmark. This framework is designed to be adaptable across various domains, with a specific focus on documents related to wind energy project permitting in this study. The contributions of this research are as follows:

**Novel domain-specific benchmark.** We present WeQA,[1] the first comprehensive benchmark QA dataset specifically designed for the wind energy domain, addressing the gap in specialized evaluation datasets for wind energy project permitting.

**Domain-agnostic framework.** Our proposed benchmark creation and LLM evaluation framework is domain-agnostic and can be tailored for any desired niche domain, enabling researchers to adapt the methodology for various specialized fields beyond wind energy.

[1]This benchmark will be made publicly available.

**Hybrid question generation.** We introduce a hybrid method that automatically generates diverse question types with varying complexity levels, producing both objective and subjective responses across different document sections to comprehensively evaluate LLM performance.

**Scalable evaluation methodology.** We utilize established scoring frameworks like RAGAS (Es et al., 2023) and incorporate multiple LLMs as judges, ensuring scalability, reproducibility, and comprehensive performance assessment of RAG-based systems.

## 2 Related Works

There have been a lot of work in the field of benchmarking, particularly for question answering (QA) task. These can be broadly divided into general QA and domain-specific QA.

**General QA benchmarks.** These benchmarks have established foundational evaluation frameworks for reading comprehension and knowledge retrieval tasks. Notable general QA benchmarks include reading comprehension datasets such as the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) and MCTest (Richardson et al., 2013), reasoning-focused benchmarks like the AI2 Reasoning Challenge (ARC) (Clark et al., 2018), and comprehensive evaluation suites such as GLUE (Wang et al., 2018) and Big Bench (Srivastava et al., 2022). Additional benchmarks targeting open-domain knowledge include CommonsenseQA (Talmor et al., 2018), TriviaQA (Joshi et al., 2017), Search QA (Dunn et al., 2017), and NewsQA (Trischler et al., 2016).

**Domain-specific QA benchmarks.** Recognizing the limitations of general benchmarks for specialized applications, researchers have developed domain-specific evaluation frameworks that capture the unique linguistic patterns, technical terminology, and reasoning requirements of particular fields. While scientific benchmarks such as MMLU (Hendrycks et al., 2020), SciBench (Wang et al., 2023), SciQ (Welbl et al., 2017), SciRepEval (Singh et al., 2022), SciQA (Auer et al., 2023), and QASA (Lee et al., 2023) are used for multi-disciplinary scientific QA evaluations, field-specific benchmarks include TheoremQA (Chen et al., 2023c) for mathematics, emrQA (Pampari et al., 2018) for medicine, BioRead (Pappas et al., 2018) and BioMRC (Pappas et al., 2020) for biology, LawBench (Chen et al., 2023b) for legal, and NuclearQA (Acharya et al., 2023) for nuclear domains.

For environmental assessment specifically, benchmarks such as EnviroExam (Huang et al., 2024) for environmental science QA and NEPAQuAD (Phan et al., 2023) for Environmental Impact Statement (EIS) documents have emerged. However, to our knowledge, no benchmarks exist specifically for wind energy project permitting, making the proposed WeQA benchmark the first comprehensive benchmarking effort in this critical domain.

## 3 Dataset Creation

In this paper, we focus on wind energy-related documents to enable the RAG-based LLMs to answer questions specific to this field. We gather PDF documents, including research articles and environmental impact studies published by the Department of Energy (DOE) under the National Environmental Policy Act (NEPA). Accessing information from this vast database is not straightforward, necessitating the need for a trained LLM to accurately retrieve and answer questions from the provided context. The challenge is to ensure that the model's responses are based on the actual documents and do not hallucinate information. By using RAG-based LLMs, we aim to enhance the reliability and accuracy of responses related to wind energy, leveraging the rich information within our extensive document collection. This approach ensures that the information provided is both relevant and grounded in the sourced material.

We constructed a data extraction and curation pipeline to extract text, image, and table information from wind energy-related documents as depicted in the 'data curation pipeline' in Figure 1. Utilizing large language model (LLM) based methods such as the *Unstructured.io* tool (Raymond, 2023), we efficiently extracted information and converted it into JSON elements. To ensure data quality, we implemented a filtering step to remove images without meaningful content, such as decorative elements or blank spaces. These filtered JSON elements were then organized into a schema, creating a page-wise assortment of text, table, and image elements. This structured format ensures that the extracted data is easily accessible and can be accurately referenced during model training and evaluation.

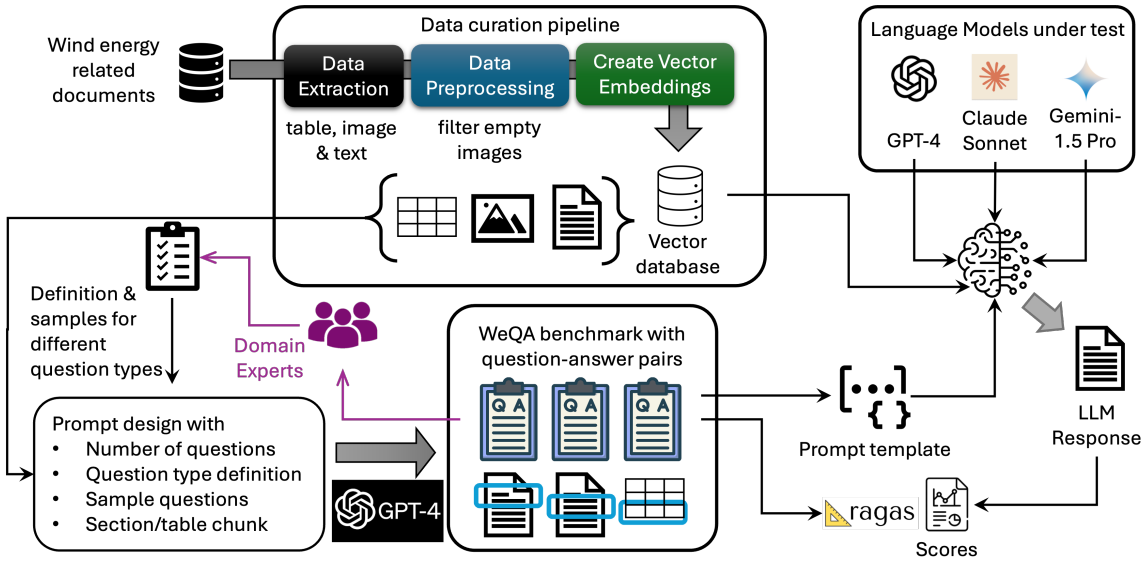


Figure 1: An overview of the proposed RAG benchmarking framework. Multiple versions of hybrid questions are generated from specific text chunks of source documents with human-in-the-loop to review them. These questions are used as prompts for the LLM or RAG model under test.

## 4 Methodology

While past works have generally preferred to use crowdsourcing as a way to craft datasets and benchmarks (Sap et al., 2019; Acharya et al., 2021), we choose to automated methods for benchmark question generation. Automatically generating benchmarking questions using GPT-4 allows for efficient and scalable evaluation of other LLMs and RAG. However, this approach can introduce errors, leading to poor quality of questions being generated. This makes it essential to incorporate a human-in-the-loop for reviewing and refining the questions and responses. This paper proposes hybrid approaches, where automated methods are combined with human curation to ensure the accuracy and reliability of the benchmarking process. By leveraging both machine and human expertise, we can achieve more robust and comprehensive benchmarking framework.

Figure 1 provides an overview of the proposed LLM benchmarking framework. The core of the benchmarking framework is the question generation aspect, where automatic generation of questions forms the foundation. We combine this with human curation to select high-quality questions, ensuring relevance and clarity. Corresponding answers to these questions are then validated by humans, establishing a reliable ground truth. This curated set of questions and validated answers is used to evaluate the responses of other LLMs and RAG models.

**Different question types.** We generate multiple types of questions, including closed, open, comparison, evaluation, recall, process, and rhetorical questions. This diversity ensures a comprehensive benchmarking process, as each question type assesses different aspects of the models' capabilities. By incorporating a wide variety of questions, we can more effectively evaluate and compare the performance of LLMs and RAG models across various dimensions. This approach provides a holistic view of their strengths and weaknesses.

Each of these question type evaluates different capabilities of the LLM under test. *Open questions* require models to generate detailed, free-form responses, testing their ability to construct coherent and informative answers. *Comparison questions* ask models to compare and contrast different concepts or entities, assessing their analytical and comparative reasoning skills. *Evaluation questions* require models to make judgments or provide assessments, gauging their ability to evaluate information critically. *Recall questions* focus on the model's ability to retrieve and reproduce specific information from memory, testing their factual accuracy. *Process questions* ask models to explain processes or sequences of actions, evaluating their understanding of procedures and logical progression. *Rhetorical questions* are used to test the models' grasp of nuances in language and their ability to recognize and appropriately respond to questions

that may not require direct answers.

We present two complementary approaches for hybrid question generation to support comprehensive LLM benchmarking. The *Hybrid Prompt Approach* employs engineered prompts to generate high-quality, curated questions, while the *Hybrid Context Approach* leverages text summarization to create questions that require broader contextual understanding. The detailed prompts used for question generation across both approaches are provided in the Appendix.

**Hybrid Prompt Approach.** We utilize GPT-4 to automatically generate questions from given text chunks through carefully designed prompts tailored to each question type. To enhance question quality, we implement a manual curation process where domain experts identify exemplary questions that effectively assess LLM capabilities for benchmarking purposes. This curation is performed systematically across all question types, ensuring that each category incorporates appropriate grammatical structures and complexity levels. These curated questions subsequently serve as few-shot examples to guide the automatic question generation framework, improving the overall quality and consistency of generated questions.

**Hybrid Context Approach.** The initial approach primarily generates questions at the sentence level by substituting subjects or objects with interrogative words, which proves adequate for 'closed', 'open', and 'recall' type questions where answers can be directly extracted from the text. However, 'process', 'evaluation', and 'comparison' questions require deeper inferential reasoning across larger text segments. To address this limitation, we first employ GPT-4 to summarize extensive text chunks (typically exceeding 15 sentences) into concise summaries containing 5-8 sentences. We then generate questions from these summarized chunks using the hybrid prompt methodology combined with curated sample questions, ensuring that the resulting questions necessitate comprehensive understanding and synthesis of broader contextual information.

**Questions from tables.** An essential component of benchmarking RAG-based LLMs within research articles and reports involves evaluating their capability to retrieve and interpret tabular information. Tables represent critical content elements within research documents, frequently containing comprehensive summaries and key quantitative data that encapsulate the essence of entire sections or studies. To address this requirement, we extract tabular data in HTML format and systematically organize it within our JSON schema framework. This HTML-formatted tabular data is subsequently incorporated into our prompt engineering pipeline to generate targeted question-answer pairs that specifically assess the model's proficiency in understanding and reasoning over structured tabular information.

Table 1: Question types in the WeQA benchmark

| Type | #Questions | % Questions |
|---|---|---|
| Closed | 382 | 18% |
| Comparison | 393 | 19% |
| Evaluation | 273 | 13% |
| Rhetorical | 324 | 16% |
| Process | 172 | 8% |
| Recall | 258 | 12% |
| Open | 270 | 13% |

Figure 2 illustrates the diverse question-answer pairs generated from the introduction section of a document (Invenergy, 2014) using our proposed methodology. We demonstrate the Hybrid Context approach where the section content is first summarized into a concise form, and subsequently, targeted QA pairs are generated from this summarized context to ensure comprehensive coverage of key concepts. Table 1 presents the statistical distribution of different question types within the WeQA benchmark, providing insights into the composition and balance of our evaluation dataset.

## 5 Results and Discussion

**Experimental setup.** We conduct a comprehensive evaluation of three state-of-the-art LLMs—GPT-4, Gemini, and Claude—on our WeQA benchmark within a RAG framework Knowledge extraction is performed from wind energy documents to create vector embeddings as shown in the data-curation pipeline in Figure 1, which are subsequently stored in a vector database to enable retrieval-augmented generation capabilities. We employ the RAGAS evaluation framework, leveraging judge LLMs to provide systematic assessment of model performance across multiple dimensions. The evaluation encompasses key metrics including answer correctness, context precision, and context recall, offering comprehensive insights into each model's proficiency in both retrieving relevant information and generating

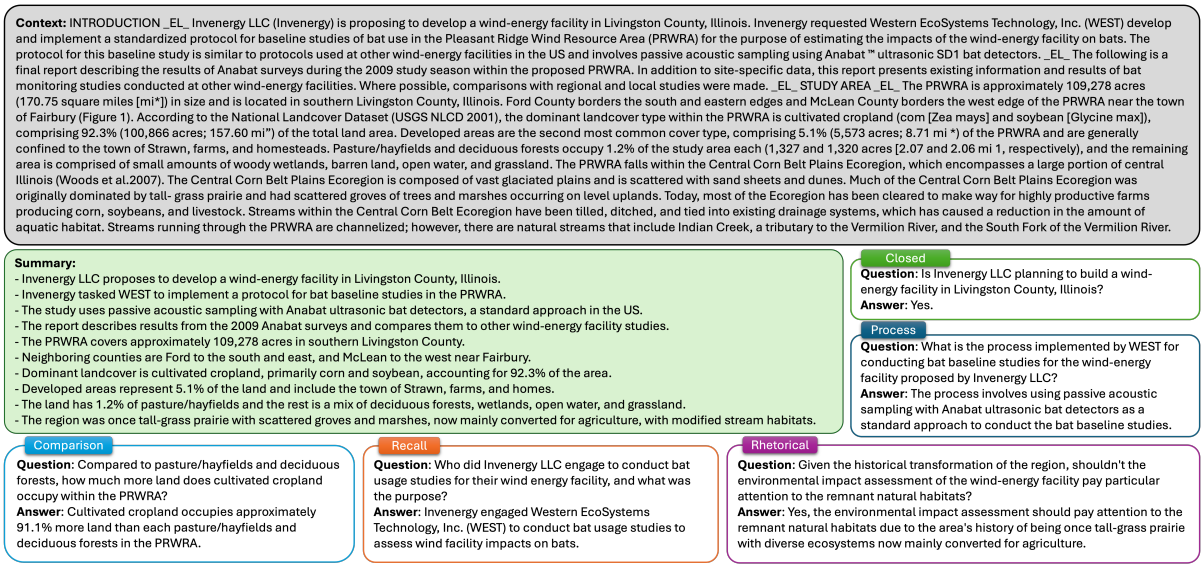


Figure 2: Different types of questions generated from the "introduction" section of a report (Invenergy, 2014) generated by the *hybrid context approach*. The section from the original document is first summarized and the question-answer pairs are generated from the summarized text chunk.

accurate responses from the provided context. For the judge LLM component, we utilize both GPT-4 and Gemini-1.5Pro to ensure robust and unbiased evaluation of the assessed models' performance. Figure 3 presents the answer correctness score, while the context precision and context recall depicted in Table 3 (added in Appendix) show the ability of the models to retrieve the context accurately.

**Observation 1** *The observed answer correctness scores are notably low, indicating a robust and challenging benchmark.*

Specifically, "evaluation" and "comparison" type questions yield nearly zero answer correctness scores for all models, highlighting their difficulty in responding. Recall that, these challenging questions were crafted from summaries of text chunks rather than the text chunks themselves, further complicating the models' ability to generate correct answers. This underscores the complexity and rigor of the benchmarking process, emphasizing the need for models to improve their understanding and contextual extraction capabilities.

**Observation 2** *There is an alignment in evaluations made by the two judge LLMs used within the* RAGAS *framework, particularly visible for 'closed' type questions.*

This similarity arises because the answers to these questions are objective ('yes' or 'no'), leading to equivalent correctness evaluations by both models. Although there are some mismatches in the evaluations made by the two judge LLMs, the number of these discrepancies is insignificant compared to the number of matching evaluations.

Figure 4 displays the confusion matrix illustrating the evaluations made by the two judge LLMs (GPT-4 and Gemini-1.5Pro) on the responses provided by the RAG-based Claude and GPT-4 models to the benchmarking questions. In this context, a true positive occurs when the judge LLM correctly identifies the model response as matching the ground truth. Conversely, a false positive arises when the judge LLM incorrectly states that the model response matches the ground truth, while it does not. This matrix helps visualize the accuracy and reliability of the evaluations conducted by the LLMs, when used within the RAGAS framework. We note that majority of evaluations made by either judge LLM matches the actual evaluation which indicates that both of them are reliable.

**Observation 3** *Comparison between 'closed' and 'open' type questions within the same section reveals a higher answer correctness for responses to 'open' type questions than 'closed' type questions.*

From this observation, we conclude that RAG-based models generate more accurate subjective responses to 'open' questions than objective ('yes' or 'no') responses for 'closed' questions. This phenomenon may stem from the inherent design of LLMs, which are optimized for generating extensive text sequences and may struggle with the precision required for definitive binary responses. This suggests that these models perform better when tasked with generating detailed, context-rich an-

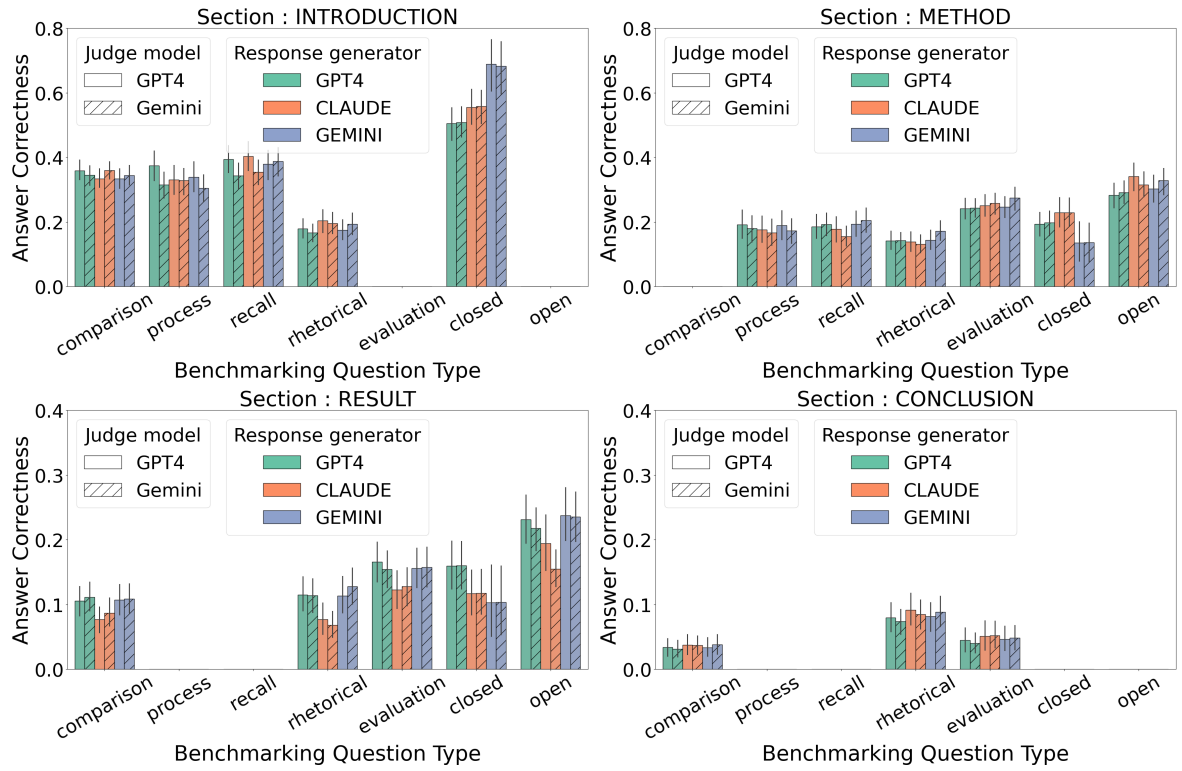


Figure 3: Answer correctness scores computed using the RAGAS scoring framework with GPT-4 and Gemini-1.5Pro as judge models for response generated by all three models used.

swers rather than simple, binary ones, highlighting their strength in handling nuanced and complex queries.

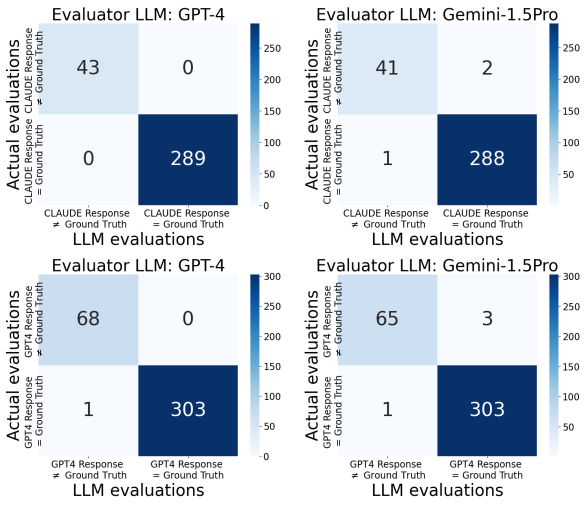


Figure 4: Confusion Matrix for evaluations by judge LLMs on responses from Claude (top) and GPT-4 (bottom) models

**Observation 4** *The answer correctness scores for questions derived from the "Introduction" section are higher compared to those from other sections.*

This is because the "introduction" section is typically longer, more similar to other documents, and often includes a related works section, which aligns closely with content found in many other documents. As a result, the RAG-based LLMs can more easily extract relevant information to answer questions accurately, leading to higher correctness scores. Additionally, the content in the "introduction" section is primarily text-based, unlike other sections which contain equations, tables, and figures. Therefore, the models provide more accurate responses to questions from the "introduction" section compared to those from other sections.

**Observation 5** *The answer correctness scores for 'rhetorical' questions are lower than those for other question types.*

This is because 'rhetorical' questions lack definite answers in the documents, making it challenging for the models to retrieve the appropriate context and provide correct responses. The absence of clear, concrete answers in the source material complicates the models' ability to generate accurate and relevant responses, leading to lower correctness scores for this question type.

**Observation 6** *Evaluations made by Gemini-1.5Pro on the responses generated by all three LLMs are higher than the evaluations made by*

*GPT-4, with the responses from Gemini LLM receiving significantly higher scores.*

Figure 5 shows the scores computed by the judge LLMs for the responses generated by the three RAG-based LLMs. The Gemini-1.5Pro judge tends to rate high scores even when the LLMs refuse to answer. An example is listed below:

**Evaluation comparison of LLM generated answer**

**Question:** In the HTML table that estimates the annual number of bird collisions at different percentages of avoidance, what is the estimated number of collisions per year at 98.0% avoidance?
**Expected answer:** The estimated number of collisions per year at 98.0% avoidance is 152.
**Gemini generated answer:** This question cannot be answered from the given source. While the text discusses bird collisions with wind turbines, it does not provide any data on estimated collision numbers at different avoidance rates.
**GPT-4 score:** 0.194
**Gemini-1.5Pro score** 0.813

Conversely, in instances where the LLMs generate correct answers, Gemini-1.5Pro has been observed to evaluate them as incorrect; such as:

**Evaluation comparison of LLM generated answer**

**Question:** Who is the GIS Technician in the 'STUDY PARTICIPANTS' table?
**Expected answer:** JR Boehrs
**Gemini generated answer:** Saif Nomani JR Boehrs was the GIS Technician.
**GPT-4 score:** 0.703
**Gemini-1.5Pro score:** 0.200

## 6 Conclusion

In conclusion, this paper presents a versatile framework for evaluating the performance of RAG-based LLMs across various question types and document sections. We showcase this by introducing a hybrid, automated question-generation method that ensures comprehensive coverage of both objective and subjective queries, and implement this for the use case of wind energy related document and present the WeQA benchmark, which is a first of its kind benchmark in wind energy domain. However, the usefulness of our work goes beyond this niche domain as our approach is domain-agnostic, meaning it can be used for creating benchmark for any domain. Additionally, our use of the RAGAS scoring framework allows for a thorough evaluation of model performance, offering a holistic assessment of LLM capabilities, while also having the advantage of being easy for other researchers to adapt this approach for their own work.

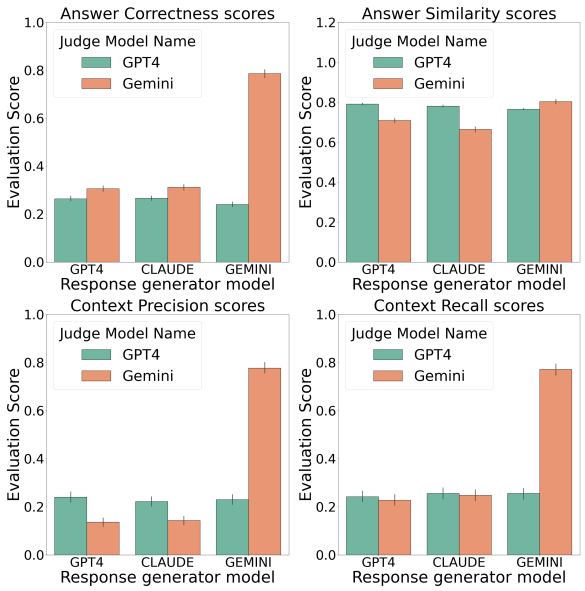


Figure 5: Answer correctness (top left), answer similarity (top right), context precision (bottom left) and recall (bottom right) scores across different judge and generator models.

## 7 Limitations

A limitation of the proposed framework is that the automatic method of generating questions often produces queries that are too specific to the document from which they were derived. When these questions are posed to an LLM with a large document corpus, the model may struggle to respond accurately, necessitating the filtering of ambiguous questions to ensure relevance and clarity. Additionally, the RAGAS scoring framework, which relies on LLMs as judges, introduces uncertainty in performance metrics, as different judge LLMs may score responses differently. While comparisons can be made for questions with objective responses, evaluating and comparing subjective responses across different LLMs remains challenging and less consistent. Another limitation of this study is the absence of comprehensive ablation studies, including comparisons between RAG-enabled and non-RAG configurations, which would provide deeper insights into the specific contributions of retrieval mechanisms to model performance.

## 8 Ethical Considerations

While we do not anticipate the novel work presented here to introduce new ethical concerns in

and by themselves, we do recognize that there may also be pre-existing concerns and issues of the data, models, and methodologies we have used for this paper. We acknowledge that researchers should not "simply assume that [...] research will have a net positive impact on the world" (Hecht et al., 2021). In particular, it has been seen that Large Language Models (LLMs), like the ones used in this work, exhibit a wide variety of bias – *e.g.,* religious, gender, race, profession, and cultural – and frequently generate answers that are incorrect, misogynistic, antisemitic, and generally toxic (Abid et al., 2021; Buolamwini and Gebru, 2018; Liang et al., 2021; Nadeem et al., 2021; Welbl et al., 2021). However, when used within the parameters of our experiments detailed in this paper, we did not see such behaviour from any of the models. To our knowledge, when used as intended, our models do not pose additional ethical concerns than any other LLM.

## References


Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Anurag Acharya, Sai Munikoti, Aaron Hellinger, Sara Smith, Sridevi Wagle, and Sameera Horawalavithana. 2023. Nuclearqa: A human-made benchmark for language models for the nuclear domain. *arXiv preprint arXiv:2310.10920*.

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2021. Towards an atlas of cultural commonsense for machine reasoning. In *Workshop on Common Sense Knowledge Graphs (CSKGs), The Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. 2023. The SciQA scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

Alan Bond, Francois Retief, Angus Morrison-Saunders, Jenny Pope, Reece C. Alberts, Claudine Roos, and Dirk Cilliers. 2024. Investigating communication of findings in environmental impact assessment and developing a research agenda for improvement. *Environmental Impact Assessment Review*, 105:107453.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. Benchmarking large language models in retrieval-augmented generation. *Preprint*, arXiv:2309.01431.

Kai Chen, D. Zhu, Jidong Ge, Zhiwei Fei, Zhuo Han, Xiaoyu Shen, Zongwen Shen, Fengzhe Zhou, and Songyang Zhang. 2023b. Lawbench: Benchmarking legal knowledge of large language models. *ArXiv*, abs/2309.16289.

Wenhu Chen, Ming Yin, Max Ku, Elaine Wan, Xueguang Ma, Jianyu Xu, Tony Xia, Xinyi Wang, and Pan Lu. 2023c. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *Preprint*, arXiv:2309.15217.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Brent Hecht, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, et al. 2021. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *arXiv preprint arXiv:2112.09544*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yu Huang, Liang Guo, Wanqian Guo, Zhe Tao, Yang Lv, Zhihao Sun, and Dongfang Zhao. 2024. Enviroexam: Benchmarking environmental science knowledge of large language models. *Preprint*, arXiv:2405.11265.

Invenergy. 2014. Bird and bat conservation strategy for Invenergy's pleasant ridge wind project.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, Enhong Chen, Yi Luo, Peng Cheng, Haiying Deng, Zhonghao Wang, and Zijia Lu. 2024. CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models. *Preprint*, arXiv:2401.17043.

Sai Munikoti, Anurag Acharya, Sridevi Wagle, and Sameera Horawalavithana. 2024a. Atlantic: Structure-aware retrieval-augmented language model for interdisciplinary science. In *Workshop on AI to Accelerate Science and Engineering, The Thirty-Eighth Annual AAAI Conference on Artificial Intelligence*, volume 3.

Sai Munikoti, Anurag Acharya, Sridevi Wagle, and Sameera Horawalavithana. 2024b. Evaluating the effectiveness of retrieval-augmented large language models in scientific document reasoning. In *Proceedings of the 4th Workshop on Scholarly Document Processing @ ACL 2024*. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. EMRQA: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Dimitris Pappas, Ion Androutsopoulos, and Harris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149.

Hung Phan, Anurag Acharya, Sarthak Chaturvedi, Shivam Sharma, Mike Parker, Dan Nally, Ali Jannesari, Karl Pazdernik, Mahantesh Halappanavar, Sai Munikoti, et al. 2023. Rag vs. long context: Examining frontier large language models for environmental review document comprehension. *arXiv preprint arXiv:2407.07321*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Partha Pratim Ray. 2023. Benchmarking, ethical alignment, and evaluation framework for conversational ai: Advancing responsible development of chatgpt. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(3):100136.

Brian Raymond. 2023. UNSTRUCTURED.IO. https://unstructured.io/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024. A survey on llm-generated text detection: Necessity, methods, and future directions. *Preprint*, arXiv:2310.14724.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *Preprint*, arXiv:2402.13178.

Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. 2024. Enhancing large language model performance to answer questions and extract information more accurately. *Preprint*, arXiv:2402.01722.

## A Prompts used to generate QA pairs using Hybrid Prompt Approach

In this section, we detail the various prompts used to create the different types of questions in the WeQA benchmark dataset. First, we show the prompt to generate questions from a given text chunk. We use curly braces to denote placeholders for the different inputs to the prompt.

**Prompt with placeholder**

Generate {**number of questions**} questions given the content provided in the following paragraph. Restrict the type of questions to {**question type**} questions.
{**Text chunk from document section**}

We curate the generated questions, where domain experts manually identify the questions which are best suited for the purpose of benchmarking LLMs. We perform this process for each type of question, so that we include particular grammatical structures for each question type. Thereafter, we use these curated high-quality questions as *few-shot examples* to regenerate questions using the automatic question generation framework. The updated prompt along with the placeholders looks as follows:

**Prompt with placeholder**

Generate {**number of questions**} questions given the content provided in the following paragraph. Restrict the type of questions to {**question type**} questions.
{**Text chunk from document section**}
You can generate similar questions (but not limited) to sample questions provided below.
{**Sample question 1**}
{**Sample question 2**}
{**Sample question 3**}

## B Prompts used to generate QA pairs using Hybrid Context Approach

We use the following prompt to summarize a document section from which the questions are to be generated.

**Prompt with placeholder**

You are a smart assistant. Can you summarize this input paragraph within {**number of points**} bullet points. Return the summarized text.
Input paragraph: {**Text chunk from document to summarize**}

Thereafter, we use the earlier prompt to generate questions from this summarized text chunk. We add the few-shot example questions which are identified by the domain experts for each question type.

**Prompt with placeholder**

Generate {**number of questions**} questions given the content provided in the following paragraph. Restrict the type of questions to {**question type**} questions.
{**Summarized text chunk from document section**}
You can generate similar questions (but not limited) to sample questions provided below.
{**Sample question 1**}
{**Sample question 2**}
{**Sample question 3**}

## C Prompts used to generate QA pairs from tables

We extract the tabular data from documents as HTML objects in the filtered JSON schema. We use the following prompt to generate question-answer pairs from the tabular data.

**Prompt with placeholder**

Generate {**number of questions**} questions given the table provided in HTML format in the following paragraph? Generate the questions keeping in mind that the caption of the table is {**Table caption obtained from document.**}
Restrict the questions such that the answers can be retrieved from the provided table in the HTML format. For each question, return 3 lines: question/ answer/ proof. Make sure there are no newline characters in the proof.
Input table:{**Table in HTML format extracted from document**}

We show an example QA pair generated from a table obtained from a document (Invenergy, 2014). Table 2 shows the table from the document for reference. An example QA-pair generated from this table is provided here.

**LLM generated question-answer pair**

**Question**: What is the acreage of Cultivated Crops within the Pleasant Ridge Project Area based on the National Land Cover Database in May of 2014?
**Answer**: The acreage of Cultivated Crops within the Pleasant Ridge Project Area is 55,946 acres.
**Proof:** The table entry under the "Habitat" column for "Cultivated Crops" corresponds with the entry under the "Acres [Hectares]" column that reads "55,946[22,641]

## D Context Recall and Context Precision

We utilize RAGAS context recall and precision metrics to evaluate the retrieval performance of our RAG-based systems, where context recall measures the proportion of relevant information successfully retrieved from the knowledge base, and context precision assesses the relevance of the retrieved context to the given query. In our setup, we employ semantic similarity-based retrieval using vector embeddings, where 'relevant context' is defined as text chunks or the document sections that contain information necessary to answer the posed questions.

Table 2: Land Cover Types, Coverage, and Composition within the Pleasant Ridge Project Area, Based on National Land Cover Database in May of 2014 (Invenergy, 2014)

| Habitat | Acres [Hectares] | % Composition |
|---|---|---|
| Cultivated Crops | 55,946[22,641] | 92.6 |
| Developed | 3,432[1,389] | 5.7 |
| Deciduous Forest | 451[183] | 0.7 |
| Hay/Pasture | 347[140] | 0.6 |
| Open Water | 122[49] | 0.2 |
| Woody Wetlands | 111[45] | 0.2 |
| Barren Land | 19[8] | 0.0 |
| Herbaceous | 3[1] | 0.0 |
| **Total** | **60,431[24,456]** | **100** |

## E Judge LLM Evaluation Analysis Through Confusion Matrices

To assess the reliability and accuracy of LLMs as judges within the RAGAS evaluation framework, we conduct a detailed analysis using confusion matrices for closed-type questions where binary ('yes'/'no') responses can be objectively compared against ground truth answers. This analysis is particularly crucial for validating the trustworthiness of automated evaluation systems in benchmarking scenarios.

**Methodology for evaluation.** We evaluate two judge LLMs—GPT-4 and Gemini-1.5Pro—by comparing their assessments of RAG-based model responses (Claude and GPT-4) against manually verified ground truth labels for closed-type questions. The confusion matrix framework allows us to quantify four key evaluation scenarios:

- **True Positive (TP)**: The judge LLM correctly identifies that the model response matches the ground truth answer.
- **False Positive (FP)**: The judge LLM incorrectly states that the model response matches the ground truth when it does not
- **True Negative (TN)**: The judge LLM correctly identifies that the model response does not match the ground truth answer

| Section ↓ | Model → / Type ↓ | GPT-4 as Judge | | | | | | Gemini 1.5 Pro as Judge | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GPT | | Claude | | Gemini | | GPT | | Claude | | Gemini | |
| | | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. | Prec. | Rec. |
| Introduction | closed | 0.467 | 0.314 | 0.500 | 0.330 | **0.570** | **0.385** | 0.392 | 0.435 | 0.424 | 0.448 | **0.467** | **0.563** |
| | comparison | 0.556 | 0.596 | **0.607** | **0.672** | 0.587 | 0.628 | 0.429 | 0.597 | **0.480** | **0.637** | 0.454 | 0.632 |
| | process | 0.565 | 0.608 | **0.598** | **0.625** | 0.586 | 0.602 | 0.457 | 0.568 | 0.467 | 0.603 | **0.483** | **0.591** |
| | recall | 0.529 | 0.597 | **0.560** | **0.617** | 0.540 | 0.586 | 0.491 | 0.611 | **0.487** | **0.624** | 0.483 | 0.601 |
| | rhetorical | 0.305 | 0.296 | **0.365** | **0.353** | 0.319 | 0.306 | 0.272 | 0.299 | **0.323** | **0.339** | 0.283 | 0.299 |
| Method | closed | 0.162 | 0.119 | **0.168** | **0.139** | 0.094 | 0.082 | 0.128 | 0.176 | **0.144** | **0.174** | 0.084 | 0.093 |
| | open | 0.364 | 0.431 | **0.431** | **0.540** | 0.378 | 0.471 | 0.333 | 0.455 | **0.383** | **0.511** | 0.367 | 0.446 |
| | evaluation | 0.400 | 0.387 | **0.442** | **0.453** | 0.416 | 0.422 | 0.311 | 0.406 | **0.352** | **0.474** | 0.316 | 0.430 |
| | process | 0.270 | 0.275 | 0.270 | 0.293 | **0.282** | **0.302** | 0.209 | 0.282 | 0.162 | 0.268 | **0.210** | **0.306** |
| | recall | 0.234 | 0.277 | 0.223 | 0.268 | **0.250** | **0.285** | **0.223** | **0.270** | 0.188 | 0.251 | 0.212 | 0.278 |
| | rhetorical | 0.229 | 0.223 | 0.241 | 0.232 | **0.250** | **0.238** | 0.208 | 0.238 | 0.193 | 0.230 | **0.224** | **0.248** |
| Results | closed | **0.143** | **0.077** | 0.102 | 0.072 | 0.076 | 0.059 | **0.120** | **0.101** | 0.093 | 0.099 | 0.070 | 0.086 |
| | open | 0.284 | 0.328 | 0.263 | 0.280 | **0.325** | **0.320** | 0.230 | 0.306 | 0.192 | 0.265 | **0.253** | **0.320** |
| | comparison | 0.167 | 0.174 | 0.139 | 0.141 | **0.172** | **0.173** | 0.128 | 0.157 | 0.098 | 0.119 | **0.134** | **0.156** |
| | evaluation | **0.272** | **0.254** | 0.217 | 0.218 | 0.257 | 0.263 | **0.226** | **0.252** | 0.171 | 0.229 | 0.209 | 0.266 |
| | rhetorical | **0.192** | **0.182** | 0.133 | 0.126 | 0.183 | 0.175 | 0.156 | 0.180 | 0.100 | 0.136 | **0.160** | **0.176** |
| Conclusion | comparison | 0.048 | 0.051 | **0.059** | **0.065** | 0.055 | 0.058 | 0.045 | 0.050 | **0.053** | **0.059** | 0.050 | 0.058 |
| | evaluation | 0.082 | 0.079 | **0.100** | **0.103** | 0.086 | 0.089 | 0.073 | 0.081 | 0.072 | 0.084 | **0.078** | **0.081** |
| | rhetorical | 0.138 | 0.141 | **0.178** | **0.171** | 0.148 | 0.147 | 0.126 | 0.148 | **0.149** | **0.165** | 0.133 | 0.144 |

Table 3: Performance of the models on the WeQA benchmark scored using the RAGAS framework across judge LLMs. The "Prec." and "Rec." mean Context Precision and Context Recall respectively, while "Type" refers to the Question Type. The best performance for each question type per judge LLM is highlighted in bold.

- **False Negative (FN)**: The judge LLM incorrectly states that the model response does not match the ground truth when it actually does

**Analysis of judge LLM performance.** The confusion matrices reveal that the majority of evaluations made by both judge LLMs align with the actual ground truth evaluations, demonstrating their reliability as automated evaluators. Specifically, both GPT-4 and Gemini-1.5Pro exhibit high accuracy rates in distinguishing correct from incorrect responses, with minimal discrepancies in their assessment capabilities.

**Cross-judge agreement.** Additionally, we observe substantial agreement between the two judge LLMs, suggesting consistency in evaluation standards across different model architectures. This cross-validation approach enhances the robustness of our evaluation methodology and provides confidence in the reliability of automated assessment within specialized domain benchmarks like WeQA.