PermitQA: A Benchmark for Retrieval Augmented Generation in Wind Siting and Permitting domain

Anonymous ACL submission

Abstract

In the rapidly evolving landscape of Natural Language Processing (NLP) and text generation, the emergence of Retrieval Augmented Generation (RAG) presents a promising avenue for improving the quality and reliability of generated text by leveraging information retrieved from user specified database. Benchmarking is essential to evaluate and compare the performance of the different RAG configurations in terms of *retriever* and *generator*, providing insights into their effectiveness, scalability, and suitability for the specific domain and applications. 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 (Large Language Model (LLM)) teaming. As a case study, we demonstrate the framework by introducing PermitQA, a first-of-its-kind benchmark on the wind siting and permitting domain which comprises of multiple scientific documents/reports related to environmental impact of wind energy projects. Our framework systematically evaluates RAG performance using diverse metrics and multiple question types with varying complexity level. We also demonstrate the performance of different models on our benchmark.

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. 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 niche scientific domain such as nuclear, renewable energy, environmental policy, etc. (Munikoti et al., 2024a,b; Phan et al., 2023)

As this innovative approach gains traction within the research community and industry applications, its effectiveness and robustness must be rigorously assessed against established benchmarks to ensure its practical utility and reliability (Chen et al., 2023a). Benchmarking is essential to establish standardized evaluation metrics and dataset that can effectively capture the nuances of text quality, coherence, factual accuracy, and relevance. Further, it facilitates comparison between RAG and existing text generation methods, shedding light on its strengths, limitations, and potential areas for improvement (Xiong et al., 2024). A robust benchmarking framework can enable researchers and practitioners to systematically investigate the impact of various parameters, such as retrieval strategies, model architectures, and training data, on the performance of RAG (Ray, 2023).

In benchmarking RAG for text generation, it is crucial to evaluate its performance across a diverse set of questions to ensure its efficacy in handling various linguistic contexts and user intents (Lyu et al., 2024). A set of well curated and diverse questions enable a comprehensive assessment of RAG's ability to generate coherent and relevant responses across various domains, ensuring its applicability in real-world scenarios. To generate such questions, automated methods leveraging NLP techniques can be employed. These methods include rule-based approaches, template filling, and neural network-based models, which enable the efficient creation of diverse question sets by leveraging linguistic patterns and semantic transformations.

Human-curated questions offer a level of linguistic richness and contextual relevance that may be challenging to achieve solely through automated generation methods (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, providing a more comprehensive evaluation of RAG's performance across diverse scenarios and applications. 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 documents related to wind turbine siting and permitting. These questions 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 in a specific domain and also proposes a generic framework to evaluate the performance of RAG-based LLMs in responding to different types of questions. This framework is designed to be adaptable across various domains, with a specific focus on wind energy-related documents in this study. The contributions of this research are as follows: **(i)** We present PermitQA, [1] the first benchmark in the Wind Siting and Permitting domain, **(ii)** our proposed framework is domain-agnostic, so it can be tailored for any desired niche domain

[1]This benchmark will be made publicly available.

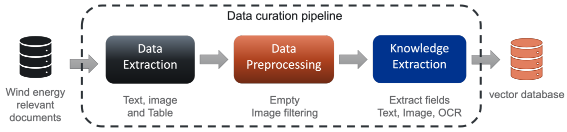

Figure 1: An overview of data extraction and curation pipeline to generate a vector database of relevant wind energy related documents.

(iii) we introduce a hybrid method to automatically generate various types of questions, producing both objective and subjective responses. The framework also generates questions from different sections of documents to evaluate LLM performance across various sections and question types, and **(iv)** we utilize existing scoring frameworks like RAGAS to evaluate model performance, incorporating different LLMs as evaluators for scoring. This approach ensures scalability and quick reproducibility of this approach, while also providing a holistic comparison of LLM performance in terms of responding to questions and assessing or comparing LLM responses with the ground truth answers.

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.

The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), consisting of 100,000+ questions and a reading comprehension dataset, is arguably the most famous general QA benchmark of the field. They contain three subtasks within QA: reading comprehension, Open-domain QA, and missing word prediction. The AI2 Reasoning Challenge (ARC) (Clark et al., 2018) is another major work, which contains almost $8,000$ science questions in English, and also included questions that neither a retrieval-based algorithm nor a word co-occurrence algorithm were able to answer correctly. Similarly, the MCTest dataset (Richardson et al., 2013) consists of 500 stories and 2000 young children level multiple-choice reading comprehension questions. Some other notable QA benchmarks include Big Bench (Srivastava et al., 2022), ARC2 (Bhakthavatsalam et al., 2021), GLUE (Wang et al., 2018), CommonsenseQA (Talmor et al., 2018), TriviaQA: 650K QA pairs with evidence (Joshi et al., 2017), Search QA (Dunn et al., 2017), NewsQA: 10K news articles (Trischler et al., 2016), *inter alia*.

More recently, there have been several benchmarks that focus on scientific and adjacent fields. The scientific portions of the MMLU (Hendrycks et al., 2020) benchmark is perhaps one of the most widely used science benchmarks, which include college and high school level questions in Physics, Chemistry, Biology, Computer Science, etc. Science Questions: 1K multiple choice questions in AI2R (Talmor et al., 2018) and SciQ Dataset: (Welbl et al., 2017) 13,679 multiple choice science questions are two other major benchmarks in the scientific domain, as is the SciQA (Auer et al., 2023), a scientific QA benchmark created by using knowledge graphs of academic articles. SciRepEval(Singh et al., 2022) is a benchmark that has four different task types – classification, regression, proximity – over scientific documents.

Similarly, some of the other most recent works include SciBench(Wang et al., 2023), a benchmark of ∼700 questions sourced from textbooks for college-level science problems and QASA (Lee et al., 2023), a QA benchmark of ∼1800 questions to test reasoning on scientific articles, specifically in AI and ML domains. There are also benchmarks that address specific fields, with TheoremQA (Chen et al., 2023b) for mathematics, emrQA (Pampari et al., 2018) for medicine, and BioRead (Pappas et al., 2018) and BioMRC (Pappas et al., 2020) for biology, and NukeBERT (Jain et al., 2020) and NuclearQA (Acharya et al., 2023) for the nuclear domain.

While these scientific domains are related to our task in terms of technological similarity, to our knowledge, there are no benchmarks for our specific field and this is the first such work. The only one that comes close is the NEPAQuAD benchmark (Phan et al., 2023) that deals with QA task for Environmental Impact Statement (EIS) documents.

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.

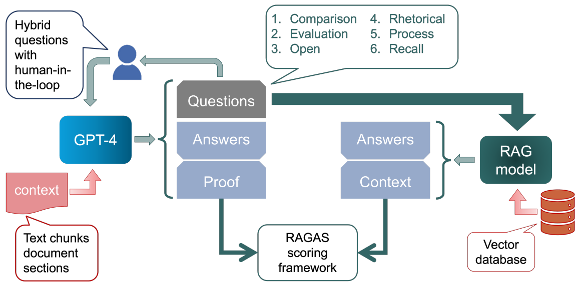

Figure 2: An overview of the proposed RAG benchmarking framework. Multiple versions of hybrid questions are generated from specific text chunks of source documents with human-in-the-loop to review them. These questions are used as prompts for the LLM or RAG model under test.

We constructed a data extraction and curation pipeline to extract text, image, and table information from wind energy-related documents as depicted 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. These 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.

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 2 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.

Next, we present two approaches for the hybrid question generation procedure required for LLM benchmarking purposes. The first approach engineers the prompt to generate well curated enhanced quality questions. The second approach summarizes the provided text chunks and generates questions from the summaries.

Summary of Introduction section made by GPT-4:
- Invenergy LLC proposes to develop a wind-energy facility in Livingston County, Illinois.
- Invenergy tasked WEST to implement a protocol for bat baseline studies in the PRWRA.
- The study uses passive acoustic sampling with Anabat ultrasonic bat detectors, a standard approach in the US.
- The report describes results from the 2009 Anabat surveys and compares them to other wind-energy facility studies.
- The PRWRA covers approximately 109,278 acres in southern Livingston County.
- Neighboring counties are Ford to the south and east, and McLean to the west near Fairbury.
- **Dominant landcover is cultivated cropland, primarily corn and soybean, accounting for 92.3% of the area.**
- **Developed areas represent 5.1% of the land and include the town of Strawn, farms, and homes.**
- The remaining land is a mix of pasture/hayfields, deciduous forests, wetlands, barren land, open water, and grassland.
- The region was once tall-grass prairie with scattered groves and marshes, now mainly converted for agriculture, with modified stream habitats.

Figure 3: Summary of "introduction" section of a report (Invenergy, 2014) generated by GPT-4. The question and the answer are generated from the summarized text chunk. The answer is retrieved from a subset of text in the chunk, shown here in red.

Hybrid prompts. We use GPT-4 to automatically generate questions from a given text chunk by providing particular *prompts* for each question type. The prompt is structured as follows:

> Generate {*num*} questions given the content provided in the following paragraph. Restrict the type of questions to {*question type*} questions.
> {*Text chunk to generate the questions.*}

An important aspect of our approach is curating the automatically generated questions to enhance the quality. To this end, we manually identify 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 identified questions as *few-shot examples* to regenerate questions using the automatic question generation framework. The updated prompt looks as follows:

> Generate {*num*} questions given the content provided in the following paragraph. Restrict the type of questions to {*question type*} questions.
> {*Text chunk to generate the questions.*}
> You can generate similar questions (but not limited) to sample questions provided below.
> {*List of sample questions*}

Hybrid text chunks. A problem with the aforementioned approach is that a significant number

Table 1: 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**

of questions are generated on a single sentence basis. This is obtained by substituting the subject or object of a sentence with a 'wh' word. These generated questions are meaningful when we consider question types such as 'closed', 'open', or 'recall', where the answers can be a single sentence from the provided text chunk. However, 'process', 'evaluation', and 'comparison' type questions of enhanced quality require the answer to be inferred from a larger portion of the given text chunk. To this end, first we use GPT-4 to summarize the entire text chunk (consisting of more than 15 sentences) into a summarized text chunk (consisting 5-8 sentences) using a prompt as follows:

> You are a smart assistant. Can you summarize this input paragraph within {*num*} bullet points. Return the summarized text.
> Input paragraph: "' {*Text chunk to summarize*} "'

Thereafter, we use GPT-4 with appropriate prompts to generate questions from these summarized text chunks using the previous hybrid prompt along with the list of sample questions. Here, we show an example question generated using this approach. We include the summary text chunk generated by GPT-4 in Figure 3 and highlight the text in red color, from which the answer for the 'comparison' type question is retrieved.

> **Question:** How does the proportion of cultivated cropland within the Pleasant Ridge Wind Resource Area (PRWRA) compared to the proportion of developed areas?
>
> **Answer:** Cultivated cropland covers 92.3% of the PRWRA while developed areas cover 5.1%.

Questions from tables. Another important aspect of benchmarking RAG models in the domain of research articles and reports is to evaluate their performance in retrieving information from tables. Tables are important contents inside research documents and often contain useful summaries of the entire documents.

> Generate {*num*} 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 are only 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*} "'

Table 1 shows a table from the report (Invenergy, 2014) and we generate questions from this table as follows.

> **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]"

5 Results and Discussion

We evaluate three RAG-based LLMs, namely GPT-4, Gemini, and Claude, on our PermitQA benchmark. The RAGAS framework is employed for this evaluation, utilizing an evaluator LLM to assess the models' performance. The assessment includes metrics such as answer correctness, context precision, and context recall, providing a comprehensive understanding of each model's capabilities in retrieving and generating accurate information from the given context. In our case, we have used GPT-4 and Gemini-1.5Pro as choices for the evaluator LLMs. Figure 4 presents the answer correctness score, while context precision and context recall depicted in Table 2 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.

		GPT-4 as Evaluator						Gemini 1.5 Pro as Evaluator					
	Model →	GPT		Claude		Gemini		GPT		Claude		Gemini	
Section ↓	Type ↓	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 2: Performance of the models on the PermitQA benchmark scored using the RAGAS framework across evaluators. 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 evaluator is highlighted in bold.

Observation 2 *There is an alignment in evaluations made by the two evaluator 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 evaluator models, the number of these discrepancies is insignificant compared to the number of matching evaluations.

Figure 5 displays the confusion matrix illustrating the evaluations made by the two evaluator 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 LLM evaluator correctly identifies the model response as matching the ground truth. Conversely, a false positive arises when the LLM evaluator 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 LLM evaluator 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 suggests that these models perform better when tasked with generating detailed, context-rich answers rather than simple, binary ones, highlighting their strength in handling nuanced and complex queries.

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

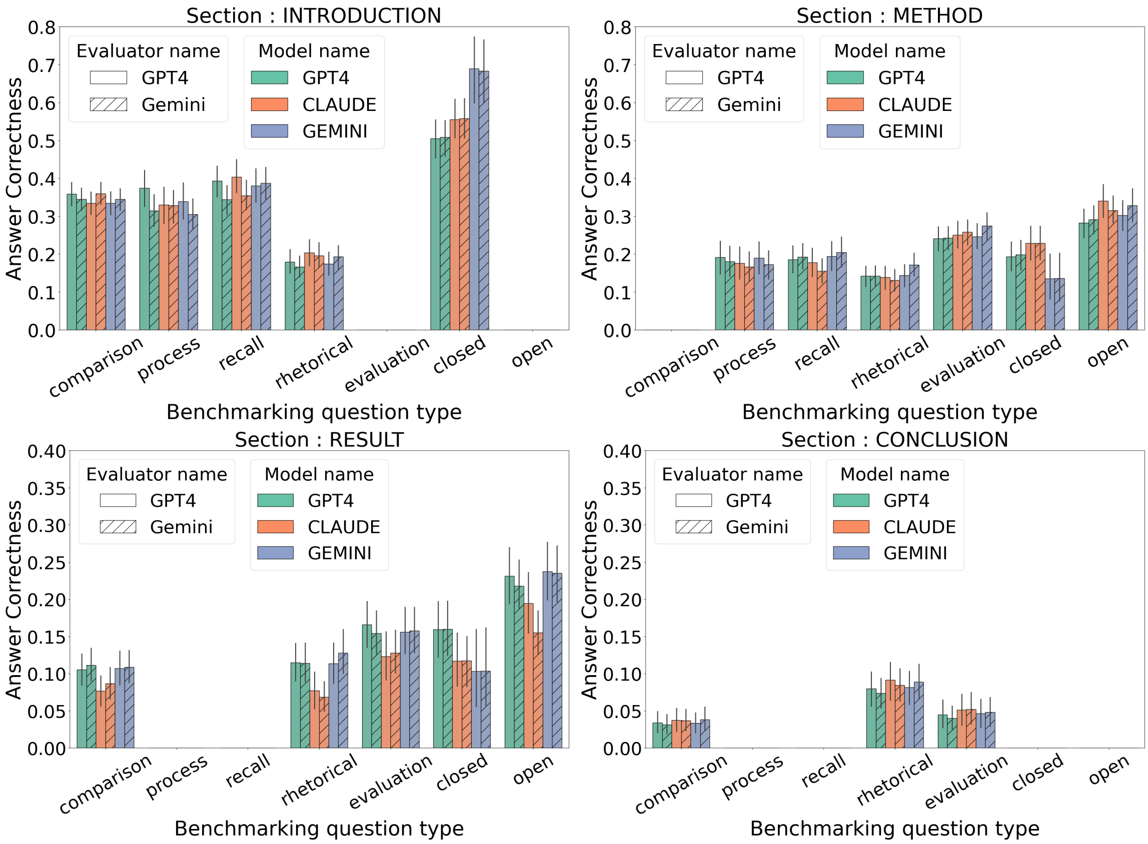

Figure 4: Answer correctness scores computed using the RAGAS scoring framework with GPT-4 and Gemini-1.5Pro as evaluator models for response generated by all three models used.

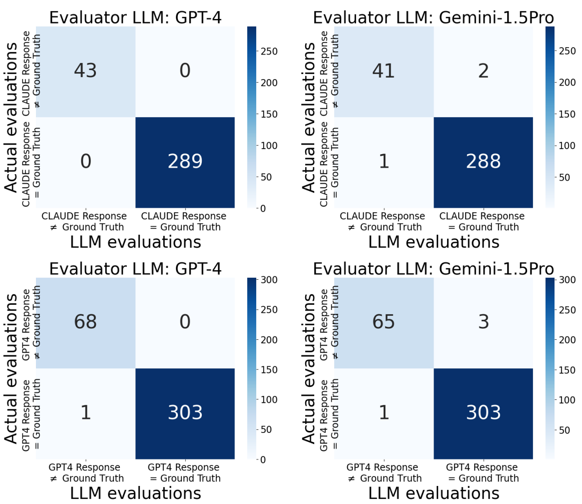

Figure 5: Confusion Matrix for evaluations by LLM evaluators on responses from Claude (top) and GPT-4 (bottom) models

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

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

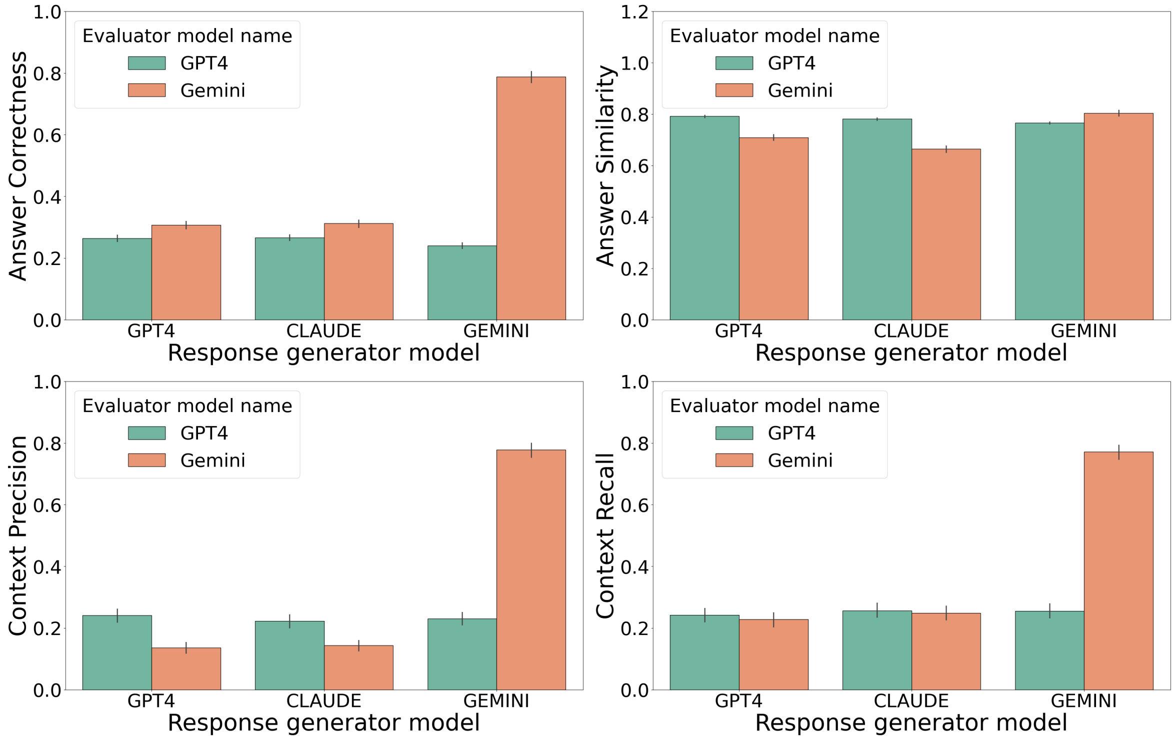

Figure 6: Answer correctness (top left), answer similarity (top right), context precision (bottom left) and recall (bottom right) scores across different evaluator and generator models.

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. An example is detailed below:

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 PermitQA benchmark, which is a first of its kind benchmark in the wind Siting and Permitting space. 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 comes with multiple benefits; it 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.

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 evaluators, introduces uncertainty in performance metrics, as different LLM evaluators 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.

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) @ 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.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? try arc-da, the direct-answer ai2 reasoning challenge. *arXiv preprint arXiv:2102.03315*.

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.

Wenhu Chen, Ming Yin, Max Ku, Elaine Wan, Xueguang Ma, Jianyu Xu, Tony Xia, Xinyi Wang, and Pan Lu. 2023b. 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*.

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*.

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

Ayush Jain, Dr NM Meenachi, and Dr B Venkatraman. 2020. Nukebert: A pre-trained language model for low resource nuclear domain. *arXiv preprint arXiv:2003.13821*.

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/`.

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*.

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*.

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.