RAG vs. Long Context: Examining Frontier Large Language Models for Environmental Review Document Comprehension

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have been applied to many research problems across various domains. One of the applications of LLMs is providing question-answering systems that cater to users from different fields. The effectiveness of LLM-based question-answering systems has already been established at an acceptable level for users posing questions in popular and public domains such as trivia and literature. However, it has not often been established in niche domains that traditionally require specialized expertise. To this end, we construct the NEPAQuAD1.0 benchmark to evaluate the performance of three frontier LLMs – Claude Sonnet, Gemini, and GPT-4 – when answering questions originating from Environmental Impact Statements prepared by U.S. federal government agencies in accordance with the National Environmental Environmental Act (NEPA). We specifically measure the ability of LLMs to understand the nuances of legal, technical, and compliance-related information present in NEPA documents in different contextual scenarios. For example, we test the LLMs' internal prior NEPA knowledge by providing questions without any context, as well as assess how LLMs synthesize the contextual information present in long NEPA documents to facilitate the question/answering task. We compare the performance of the long context LLMs and RAG powered models in handling different types of questions (e.g., problem-solving, divergent). Our results suggest that RAG powered models significantly outperform the long context models in the answer accuracy regardless of the choice of the frontier LLM. Our further analysis reveals that many models perform better answering closed questions than divergent and problem-solving questions.

1 Introduction

As Large Language Models (LLMs) become increasingly commonplace, researchers have discovered that these models are useful for many tasks beyond text generation. Specifically, LLMs have shown potential utility in niche domains (like science) that would traditionally require specialized expertise, both in a pure text setting (Horawalavithana et al., 2022; Munikoti et al., 2024) and by incorporating data of various modalities (Dollar et al., 2022; Horawalavithana et al., 2023). Recent work has been done to evaluate these models (Acharya et al., 2023; Munikoti et al., 2023; Cai et al., 2024) and to assess their uncertainty (Wagle et al., 2024). Despite extensive research, constructing LLMs for answering domain-specific questions has proven challenging (Kasneci et al., 2023).

One such challenge for LLM-based question-answering systems occurs when systems are tasked with surfacing answers to questions from the content of long documents in specialized domains. Existing LLMs allow users to include a paragraph as context along with the content of the question; however, LLMs generally limit the size of that paragraph to a specific number of tokens. This restriction forces users to truncate or manually summarize the content of lengthy documents into short passages. Another approach users can take includes submitting only the question and relying on the models to find the correct document from a corpus and relevant content needed to answer the question. This strategy often works well for answering questions from well-known domains (e.g., sports or education); however, it is not as successful for less pervasive topics (Munikoti et al., 2023). Because LLMs are data-driven, they are not as apt to provide accurate answers for questions about less popular, more specialized domains such as Law (Kapoor et al., 2024)) and Energy (Buster et al., 2024).

In this work, we focus on assessing the long-context LLMs in the environmental reviews conducted under the National Environmental Policy Act (NEPA)[1]. NEPA is a U.S. environmental law

[1] https://www.epa.gov/nepa

designed to protect the environment. An environmental impact statement (EIS) is required by Section 102(2)(C) of NEPA for any proposed major federal action significantly affecting the quality of the environment. An EIS is a detailed document that describes a proposed action, alternatives to the proposed action, and potential effects of the proposed action and alternatives on the environment. An EIS contains information about environmental permitting and policy decisions and considers a range of reasonable alternatives, analyzes the potential impacts resulting from the proposed action and alternatives, and demonstrates compliance with other applicable environmental laws and executive orders.

Along with the fact that EIS documents are usually lengthy (often several hundred pages) and are created by NEPA experts, another factor that can hinder the application of LLMs in this domain is that the development of an EIS document requires NEPA experts with various subject matter expertise to engage in preparation over multiple years, often citing older articles from as far back as the 1990s. For example, the Executive Order (EO) 12898, issued in 1994, is cited on page 60 of the EIS documents prepared for the First Responder Network Authority project[2]. This could present significant challenges for current LLMs in helping NEPA users automatically retrieve answers from LLM-based question-answering systems. To our knowledge, there is no ground-truth benchmark built specifically for this domain to evaluate the quality of LLMs' output for QA task when the questions pertain to EIS documents.

In this work, we leverage both long context and Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) to develop LLM capability for question-answering over EIS documents (Figure 1). We select frontier LLMs for our experiments: Claude Sonnet (Team and Collaborators), Gemini (Team and Collaborators, 2024), and GPT-4 (OpenAI, 2024). To assess the efficacy of our proposed RAG model compared to other context-augmentation strategies, we also conduct rigorous experiments evaluating LLMs with various types of contexts for NEPA documents. To evaluate our approach, we establish a benchmark using ground truth triples of questions, answers, and corresponding contexts, generated through a semi-supervised

[2]https://www.energy.gov/nepa/eis-0530-nationwide-public-safety-broadband-network-programmatic-environmental-impact

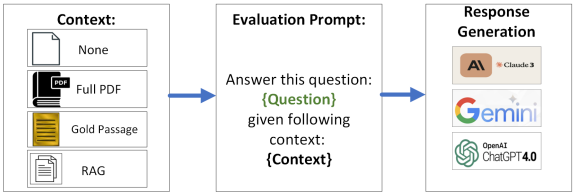


Figure 1: Illustration of varied EIS contexts used in the comparison.

method employing GPT-4. Overall, we make the following contributions:

1. Created the first-ever preliminary benchmark (NEPAQuAD1.0) to automatically evaluate the performance of LLMs in a question-answering task for EIS documents

2. Evaluated the capability of LLMs in question-answering tasks over long documents

3. Conducted rigorous comparisons of LLMs using zero-shot prompting versus context-driven prompting (i.e., passage, PDF, and RAG) to assess model performance.

The structure of this paper is outlined as follows: In Section 2, we describe the benchmark creation to assess the quality of our model in comparison to models derived from different contexts. The Section 3 section lays out our approach and the various contexts used for evaluation implementation, followed by a detailed analysis of our performance in Sections 4. Section 5 then discusses other works in literature that deal with long context and RAG for long documents. We finish with the conclusion and limitations of our work in Sections 6 and 7.

2 NEPAQuAD Benchmark

In this section, we describe the construction of a ground-truth benchmark to evaluate the quality of automated responses generated from LLMs. Due to the high costs associated with manually creating human-written questions and answers, and the inability to use ground-truth benchmarks from other domains, we adopt a semi-automatic approach to generate the NEPAQuAD1.0 (**N**ational **E**nvironmental **P**olicy **A**ct **Qu**estion and **A**nswering **D**ataset) benchmark. The general idea of our evaluation benchmark generation process is to extract meaningful passages from a set of EIS documents, then use GPT-4 to generate questions based on these passages. To ensure the quality of

the generated benchmark, two authors of this study, who are subject matter experts in NEPA, measure the quality of the ground-truth answers by comparing the provided proofs against the original context from which the questions were derived. Our generated ground-truth benchmark is a set of triples containing a question, an answer, and the proof (i.e., the text directly related to the answer, derived from the context from which the question originated). The process of benchmark generation is illustrated in Figure 2.

To evaluate performance of LLMs for the EIS question-answering task, we first select high-quality documents from the EIS document database and extract paragraphs as context to be used in the evaluation. Then, we identify the types of questions that we want to use to evaluate the models. Next, we use GPT-4 (OpenAI, 2024) to generate question-answer pairs based on the selected contexts by using carefully designed prompts. Finally, we use these generated questions to evaluate different LLMs with various contexts, with the generated answers acting as the ground-truth. We describe the process in detail below.

2.1 Gold Passage Selection

Document Selection NEPA experts select nine EIS documents from different government agencies that were most representative of various NEPA actions. These document exhibit great variations in content and focus depending on the authoring government agency, as each agency may interpret and implement the NEPA guidelines distinctively. For instance, the U.S. Forest Service might emphasize forest management and wildfire mitigation, while the Army Corps of Engineers could prioritize water resource development and infrastructure impacts. Table 3 shows the statistics about the selected documents (see Appendix). Each document has around 400 pages on average while the longest document contain more than 600K tokens.

Excerpt Selection For each of nine selected documents, we attempt to select excerpts that have important content of each document. Again, the default approach of randomly extracting excerpts poses the risk of resulting in *low-quality* excerpts, such as parts of appendix or images' captions. To avoid this risk, NEPA experts manually select excerpts from the documents. They divided each document into three sections: beginning, middle, and end, and then selected two, six, and six excerpts from each of these sections respectively, for a total of 10 excerpts from each document. We use these excerpts as the ground-truth context, called gold passages, for question benchmark generation.

2.2 Question Type Selection

Once we identified the gold passages, NEPA experts select the type of questions to include in the benchmark. We started with a list of 15 types of questions[3], and eventually narrowed it down to 10 types of questions after extensive discussions. These types are shown in Table 1. In addition to selecting the question types, the NEPA experts also created sample questions for each category for the EIS document domain. For a more detailed description of the question types, as well as example questions, please see Appendices A and B.

Question Type	#Questions	
Closed	789	49%
Comparison	64	4%
Convergent	109	7%
Divergent	121	8%
Evaluation	64	4%
Funnel	127	8%
Inference	101	6%
Problem-solving	11	1%
Process	108	7%
Recall	105	7%
Total	1599	100%

Table 1: Statistics on question types used in the NEPAQuAD1.0 benchmark

2.3 Prompt Design

The next step is to design a prompt that can instruct the question generation model to create high quality questions and answers. To ensure that the prompt can instruct generative model efficiently, we took advice from the NEPA experts to create the prompts. We also use the sample questions created by the NEPA experts to augment the original prompts and create an "enhanced" prompt. The template for the prompt and benchmark creation process is displayed in Figure 2. We restricted the output for each prompt in a CSV format with three fields: question, answer, and proof. The "proof" column stored the part of the gold passage that the model picked as evidence for the provided answer to the question.

[3]https://tinyurl.com/3akej8ct

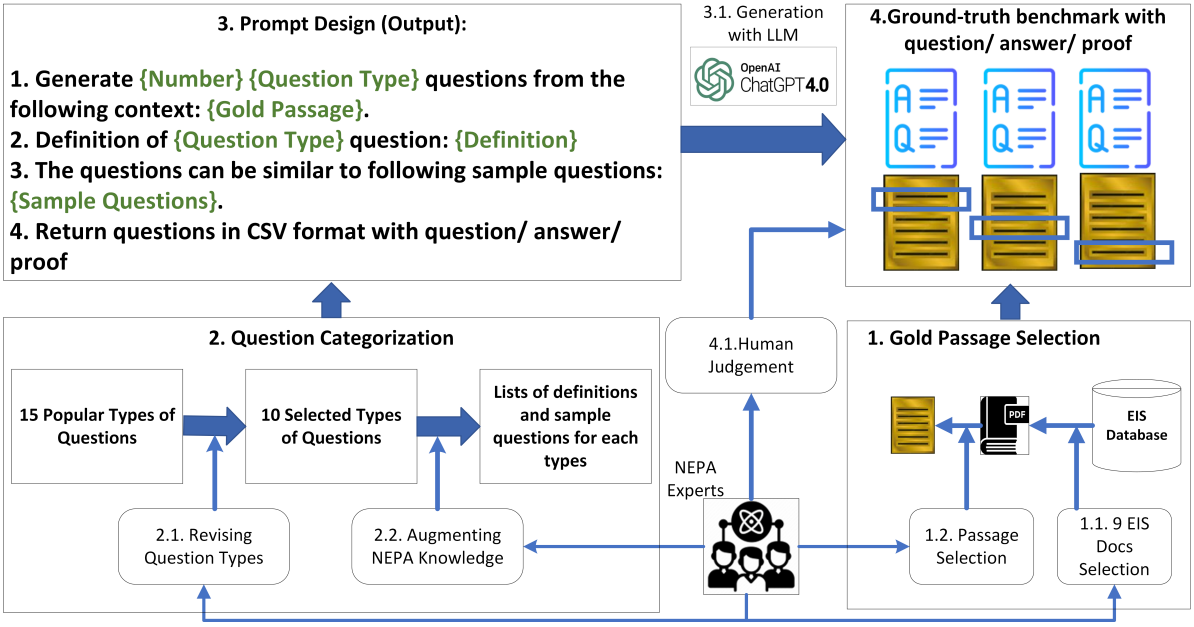


Figure 2: Steps of Ground-truth benchmark generation for evaluating LLMs over varied contexts for question-answering over EIS documents

2.4 Ground-truth Benchmark Generation

Automated Generation We selected GPT-4 as the generative model for this task, as GPT-4 has been used for generating questions and answers for documents in other domains such as agriculture (Balaguer et al., 2024). Specifically, we use GPT-4 Azure OpenAI service with default setting to execute the question generation prompts. For each document in our nine selected documents, we have 10 gold passages, results in 90 gold passages in total. For each gold passage, we generated 10 sets of questions for each of the 10 question types. We then filtered the generations for incorrect formats. Overall, we generated a benchmark of 1599 ground-truth triples of question-answer-proof over the nine selected EIS documents.

Quality Check In order to judge the generated benchmark's correctness, we randomly select 100 sample questions for validation. The validation was done by the NEPA experts (two co-authors). Each of them independently went through the sample questions, checking both the correctness of question type (i.e. whether the generated question was the same type of question as requested) and the correctness of the answer. For each question, if either of the evaluators marked the question or answer as correct, we labelled them as correct. Overall, our generated benchmark achieved 77% of answer correctness and 82% of question type correctness.

3 Experimental Setup

In this work, we experimented with three frontier LLMs: Claude-3 Sonnet, Gemini, and GPT-4. For the context provided to the model, we had four possible variations: no context, PDF documents, silver passages (RAG setup), or gold passages. The combination of the models and context settings resulted in a total of 12 unique configurations (Figure 1). We explain these configurations in details in Section 3.1 and report the evaluation metrics in Section 3.2.

3.1 Context Variation

No Context: In this setting, we simply query the models with the questions with no context provided, the same as in other general domains. We do not provide any additional context about the origin of the questions, so the models were expected to answer questions from their existing knowledge. While this strategy can work well with popular domains such as sport or literature, we assume that NEPA domain may be challenging for the LLM models to get the accurate answer. In the other word, this setting can be said to be a test of the LLMs to answer out-of-general-domain questions. As such, this setting is usually expected to return low performance.

Full PDF as Context: In this scenario, in addition to the question, we also provide the model the

PDF (text) document from which the context to the question was extracted. Since we do not inform the model which part of the document to look at, the generated responses' accuracy will heavily depend on the models' ability to pick the correct context from the very large scale textual information provided. We expect this setting to yield performance better than no context.

RAG Context (Silver Passages): In RAG models, when a question is inputted for LLM generation, the corresponding context is extracted as a relevant passage from a given EIS document. We use the standard RAG setup where we encode both question and retrieved passages with BGE embedding model (Xiao et al., 2023). We use the cosine similarity score to assess the similarity between the question and the contexts. The number of top-ranked relevant passages extracted, referred to as top-K silver passages, is set at $k = 3$.

Gold Passage as Context: In this configuration, we include the actual context from which the question was generated in the prompt, alongside the question content. While the scenario where users manually identify the correct passage is rare in practice, we simulate this scenario to measure how well LLMs can perform if we were able to extract relevant passages with very high accuracy. We expect this setting to perform the best.

3.2 Evaluation Metrics

In order to evaluate the performance of the models in different configurations, We compare the answers from the generated responses of the model across these various configurations. Overlap based metrics such as BLEU score (Papineni et al., 2002), while used by many prior works, simply measure the syntactic similarity, and as such is not suitable to perform evaluations where semantics is more important. As such, for our work, we use the RAGAs score proposed by Es et al. (2023): the Answer Correctness (called Correctness in this paper).

The Answer Correctness score combined two aspects of factual correctness and semantic correctness for its calculation. While factual correctness captured the correctness at phrase/clause level of input answer, the semantic score is achieved by comparing the similarity between vectors of expected answers and predicted answers by using embedding models. GPT-4 is used in calculating the answer correctness that quantifies the factual overlap between the generated answer and the ground truth answer (Es et al., 2023). We use the BGE (Xiao et al., 2023) as the embedding model for semantic correctness calculation. We set the weight of factual correctness as 0.75 and the weight of semantic correctness as 0.25 for measuring the Answer Correctness.

4 Performance Analysis

In this section, we describe the overall performance of the LLMs in the question/answering task evaluated with NEPAQuAD1.0 (as presented in Section 2). First, we compare the performance of three frontier LLMs: Claude-3 Sonnet, Gemini, and GPT-4 across various QA contexts (Section 4.1). Second, we compare the model performance across various question types (Section 4.2). In Section 4.3, we evaluate how the models performing to the questions generated from different parts of the document. Finally, we analyze the performance.

4.1 Evaluating Different QA Contexts

Table 2 reports the overall performance of the models across various QA contexts used in the evaluation. We observed that for the task with no context, the Gemini model produces the most accurate results by far. However, when PDF documents are provided as context, this trend is reversed, with GPT-4 surpassing Gemini in correctness. Despite that Gemini is able to handle very long contexts (1.5M tokens), it is surprising to see its performance drop when provided with PDF documents as additional contexts. This may be due to the model struggling to reason over the large amount of relevant and irrelevant content in the EIS document.

Overall, RAG models perform better in comparison to the models provided with PDF documents as additional contexts. In RAG setup, The Claude model outperforms both other models in term of correctness, although the scores across the Claude and GPT-4 models are much closer. There is notable increase in Gemini's performance in the RAG setup when compared with the PDF contexts.

As expected, all models perform best on average when provided with the gold passage in comparison to other context variations. In this scenario, model needs to synthesize information that directly contains the answer to the question posed to the model. Notably, models perform comparably when they are provided with the RAG and gold passage contexts.

Context	Claude	Gemini	GPT-4
None	21.50%	50.16%	20.28%
Complete PDF	23.47%	46.62%	56.40%
Silver Passages	68.74%	57.06%	66.86%
Gold Passage	68.41%	61.81%	67.66%

Table 2: Evaluation on the answer correctness of LLMs over different configurations of context over EIS documents. Silver passages are selected by the RAG model.

4.2 Evaluating Different Question Types

We analyze the performance of LLMs over different types of questions depicted in Figure 3. When analyzing the results based on the type of questions, we see that all three models have superior performance on closed questions when provided with either silver or gold data as context, while GPT-4 is the only model that performs well on these questions even when provided with just the PDFs as context. For all other categories, both Claude and GPT-4 exhibit similar behavior pattern when provided with none or PDF context, although GPT4's performance is notably better than Claude's in almost every category, with this difference particularly noticeable with PDF context.

Interestingly, Gemini performs really well when provided with no context at all, and the performance decreases for all categories except the convergent and recall questions when provided with the PDFs as context. For a majority of the question types, even the silver or gold data is not able to get the performance to the level of no-context, with the closed and recall questions being the exceptions. Overall, RAG models and models with other contexts performed best in answering closed questions and worst in answering divergent and problem-solving questions.

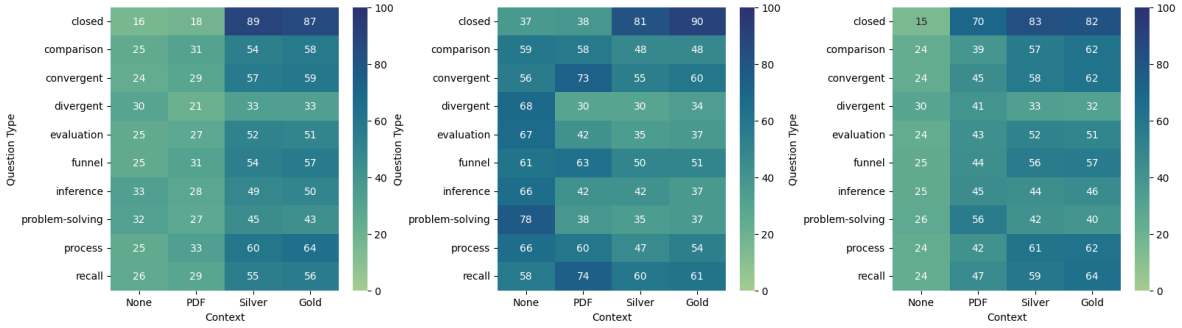
4.3 Evaluating Positional Knowledge

We also analyzed the performance of various question types based on the portion of the document from which they were derived, as shown in Figure 4. Across the models, we observed a general trend where the earlier the source text in the document, the better the performance of the models. A notable exception to this are the problem-solving questions, which perform better when sourced from the latter parts of the document. This pattern holds true for all three models. Additionally, we noticed that divergent questions yield better results when derived from the middle of the document. Overall, all three models exhibit similar or comparable patterns of performance across different document sections and question types. These results suggest that performance of long-context models may vary not only by the position of relevant information, but also due to the type of the question and the amount of reasoning that the model has to perform.

4.4 Discussion

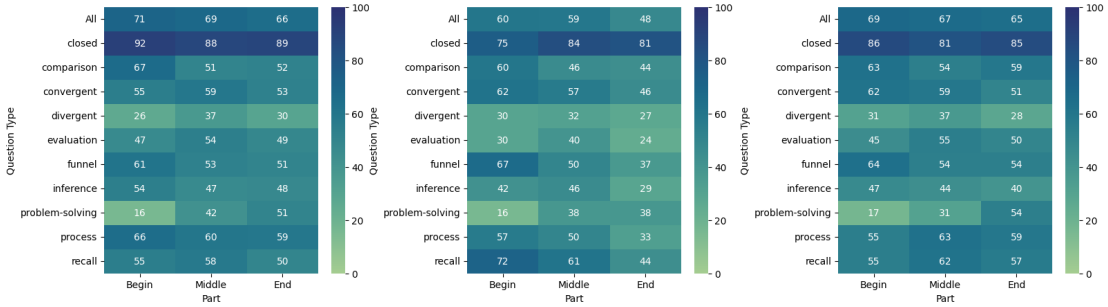
RAG for Question-Answering Over Domain-Specific Documents The findings of this study underscore the significant role of RAG models as crucial strategies for addressing domain-specific questions. These models have shown remarkable superiority in performance compared to zero-shot knowledge and using the full PDF as context. While evaluating LLMs in well-established fields such as mathematics or biology can be straightforward, using numerous human-written sets of ground-truth questions and answers (Team and Collaborators, 2024), evaluating domain-specific LLMs necessitates unsupervised or semi-supervised approaches to generate evaluation benchmarks. We recognize that while our approach satisfies the need for an automated method in this domain, it still faces challenges, particularly that the selected question types might not be representative of other research areas. Therefore, researchers in other domains should carefully consider the types of questions they want to generate for their studies.

Patterns of Output From this study, we draw two overall conclusions regarding the output patterns. First, surprisingly, we did not observe specific patterns of correctness in relation to document metadata such as token counts. This finding contradicts our initial assumption that documents with the lowest token counts would achieve the lowest accuracy and vice versa. We believe one reason for this may be that we selected only 90 passages as gold passages from the document, which might not be representative. Second, we noted that each LLM model tends to have distinctive response types. For instance, Gemini's responses tend to be straightforward when no context is provided, often stating "I don't have the context." In contrast, Claude and GPT-4 are more likely to attempt clarifications of input questions, such as predicting and providing the full content of abbreviations. We encourage researchers in other projects involving RAG to analyze patterns of output to enhance their work.



(a) Claude (b) Gemini (c) GPT-4

Figure 3: The evaluation results meassured by the Answer Correctness scores of each LLM used with 4 scenarios of using context over each question types



(a) Claude (b) Gemini (c) GPT-4

Figure 4: The evaluation results measured by the Answer Correctness scores of each LLM used over different parts with silver passages provided as context

Long Context Reasoning One of the primary objectives of this study is to assess how beneficial the long context models that can process 128K to 1.5M tokens context in answering questions from long EIS documents. We noticed that these models struggle to use long input contexts to answer more difficult questions that require multiple steps of reasoning (e.g, problem-solving). Given that we see the model performance varies over the positions and types of questions, we assume that effective question-type and -complexity aware reranking of retrieved documents may help to improve the performance (Jeong et al., 2024). For example, we can use of another LLM to adjust the order of retrieved chunks based on the problem-solving question type.

5 Related Work

Long Context Evaluation A popular technique to evaluate the long context LLMs is with a simple *needle in a haystack* analysis to test in-context retrieval ability (Chandrayan et al., 2024). Despite the claims made in these tests, it is shown current LLMs perform poorly in processing and understanding long, context-rich sequences in rigorous scientific benchmarks (Li et al., 2024; Liu et al., 2024). For example, Li et al. (2024) constructed LongICLBench benchmark to assess a set of long-context LLMs in an extreme-label classification task as a long in-context learning task. They showed that long context understanding and reasoning is still a challenging task for the existing LLMs. Few studies showed that long-context LLMs are affected by the position of the relevant information in the input context (Liu et al., 2024; Ivgi et al., 2023). For instance, Ivgi et al. (2023) showed that encoder-decoder models have significantly higher performance when relevant information is placed at the start of the input context. In addition, Liu et al. (2024) showed that LLMs perform weakly when they must access relevant information in the middle of long contexts.

There are some other benchmarks proposed in multiple languages and domains for evaluating LLM's long context understanding. Bai et al. (2023) proposed LongBench that covers six tasks, single-doc QA, multi-doc QA, summarization, few-shot learning, synthetic tasks, and code completion in English and Chinese languages. L-Eval Bench-

mark (An et al., 2023) contains 20 sub-tasks, 508 long documents, and over 2,000 human-labeled query-response pairs with diverse question styles, domains, and input length. Li et al. (2023) proposed LooGLE that includes around 6,000 questions across diverse domains and evaluated both commercial and open-sourced models. While they showed that commercial models outperform open-sourced models in short question-answering and cloze tasks they struggled in long dependency tasks. Furthermore, retrieval-based techniques showed significant advantages for answering short questions, whereas methods aimed at increasing the length of the context window had a minimal effect on the comprehension of longer contexts. ∞Bench (Zhang et al., 2024) consists of 12 tasks with data length surpassing 100K tokens on average. They suggested that long context LLMs still require significant advancements to effectively process 100K+ context.

RAG for Long Documents RAG models offer a promising approach for enabling LLMs to search and extract relevant information from lengthy documents or extensive collections. The common strategy of splitting documents into smaller, more manageable chunks that fit within the LLM's context window has its limitations, as highlighted by recent studies (Barnett et al., 2024). These limitations include the model's failure to accurately extract answers even when they are present within the provided context, particularly when there is excessive noise or contradictory information. To address these challenges, researchers have proposed novel techniques. HippoRAG, a neurobiologically inspired long-term memory system designed for LLMs to handle long documents more effectively, aims to mitigate the limitations of current RAG models (Gutiérrez et al., 2024). Gao et al. (2023) provide a comprehensive survey of RAG methods, categorizing them into three paradigms: Naive RAG, Advanced RAG, and Modular RAG. The authors highlight the remarkable adaptability of Modular RAG, which allows for module substitution or reconfiguration to address specific challenges, surpassing the fixed structures of Naive and Advanced RAG. Modular RAG integrates new modules or adjusts interaction flow among existing ones, enhancing its applicability across different tasks. The survey also discusses the concept of adaptive retrieval in RAG, exemplified by methods like Flare (Jiang et al., 2023) and Self-RAG (Asai et al., 2023). These approaches refine the RAG framework by enabling LLMs to actively determine the optimal moments and content for retrieval, enhancing the efficiency and relevance of the sourced information. Despite these advancements, Gao et al. (2023) emphasize that further research is needed to fully understand the intricacies of applying RAG to long documents and to develop more robust and reliable methods.

6 Conclusion

In this study, we conduct the initial investigation into the performance of LLMs within the domain of the National Environmental Policy Act and its associated documents. To facilitate this, we introduce NEPAQuAD, a question-and-answering benchmark designed to evaluate a model's capability to understand the legal, technical, and compliance-related content found in NEPA documents. We assess three frontier LLMs designed for handling extensive contexts—Claude Sonnet, Gemini, and GPT-4—across various contextual settings. Our comprehensive analysis indicates that NEPA documents pose a significant challenge for LLMs, particularly in terms of understanding the complex semantics and effectively processing the lengthy documents. The findings reveal that models augmented with the RAG technique surpass those LLMs that are simply provided with the PDF content as long context. This suggests that incorporating more relevant knowledge retrieval processes can significantly enhance the performance of LLMs on complex document comprehension tasks like those found in the NEPA domain. In addition, we noticed that these LLMs struggle to use long input contexts to answer more difficult questions that require multiple steps of reasoning. For example, models performed best in answering closed questions and worst in answering divergent and problem-solving questions.

7 Limitations

Similar to other applications of LLMs, our proposed system for EIS long documents also has some limitations. We list those limitations as follows:

Restriction of token limitation on full PDF context. While we are able to use the Gemini model with token length as 1.5 million, we could only use 128K tokens per query for response generation with Claude and GPT-4. Thus, we need to

truncate the content of Full PDF to run these two LLM models. This might cause the performance drop on the context as Full PDF with questions from EIS documents. In future work, we should analyze more carefully about the impact of token truncation for Full PDF context.

Uncertainty of generated responses by LLMs. Due to budget constraints, we conducted only one phase of response generation across different configurations. This introduces a risk of uncertain outputs, meaning that LLMs might generate different responses each time, even with the same input, as demonstrated in another study (Wagle et al., 2024). In future work, we plan to run LLMs multiple timee and analyze the effect of this uncertainty in response generation.

Challenges of human judgment. Currently, we leverage human evaluation as a preliminary proxy measure for qualitative analysis of benchmark. In future work, we plan to involve more NEPA experts in a more systematic manner to expand the dataset with human judgment results and to perform proper adjudication meetings between NEPA experts to reconcile conflicting results.

Bias in automated evaluation There might be a potential bias in the answer correctness evaluation process due to the use of GPT-4 to assess the outputs of various models. There is a concern that GPT-4 may inherently prefer the outputs generated by the same model over others in the factual correctness evaluation. This could lead to skewed evaluation results, where GPT-4's outputs are rated more favorably, not necessarily because they are superior, but because of the inherent biases in the evaluation model (GPT-4).

To address the potential bias in the answer correctness evaluation process, we assess both factual and semantic correctness in the evaluation. For semantic correctness, we utilize the BGE (Xiao et al., 2023) as the embedding model and we calculate the semantic similarity between the model's outputs and the ground-truth answers independently of GPT-4's own evaluation mechanisms. By combining both factual and semantic correctness, we aim to accurately reflect the true performance of various models, including GPT-4.

8 Ethical Consideration

It has generally been the norm to assume that previously published work can be used as-is without having to consider the inherited ethical issues. However, in present times, researchers should not "simply assume that [...] research will have a net positive impact on the world" (Hecht et al., 2021). We acknowledge that this applies not just to new work, but also when using existing work in the way that we have done.

While working on this project, care has been taken to ensure that any and all data was anonymized and no Personally Identifiable Information (PII) is present in the data used. We had domain experts in the team throughout the process, thereby ensuring they were aware of all the potential risks and benefits.

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. In particular, it has been seen that LLMs, like the ones used in this work, exhibit a wide variety of bias – 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*.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *Preprint*, arXiv:2401.08406.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. *arXiv preprint arXiv:2401.05856*.

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.

Grant Buster, Pavlo Pinchuk, Jacob Barrons, Ryan McKeever, Aaron Levine, and Anthony Lopez. 2024. Supporting energy policy research with large language models. *arXiv preprint arXiv:2403.12924*.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Yongge Li, Mujie Lin, Shuwen Yang, et al. 2024. Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint arXiv:2403.01976*.

Kedar Chandrayan, Lance Martin, gkamradt, Lazaro Hurtado, arkadyark cohere, Ikko Eltociear Ashimine, Pavel Král, and Prabha Arivalagan. 2024. *gkamradt/LLMTest$_N$eedleInAHaystack*.

Orion Walker Dollar, Sameera Horawalavithana, Scott Vasquez, W James Pfaendtner, and Svitlana Volkova. 2022. Moljet: multimodal joint embedding transformer for conditional de novo molecular design and multi-property optimization.

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, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *arXiv preprint arXiv:2405.14831*.

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*.

Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott Vasquez, Robin Cosbey, Maria Glenski, and Svitlana Volkova. 2022. Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 160–172.

Sameera Horawalavithana, Sai Munikoti, Ian Stewart, and Henry Kvinge. 2023. Scitune: Aligning large language models with scientific multimodal instructions. *arXiv preprint arXiv:2307.01139*.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *Preprint*, arXiv:2305.06983.

Sayash Kapoor, Peter Henderson, and Arvind Narayanan. 2024. Promises and pitfalls of artificial intelligence for legal applications. *arXiv preprint arXiv:2402.01656*.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

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.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language

models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

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.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Sai Munikoti, Anurag Acharya, Sridevi Wagle, and Sameera Horawalavithana. 2023. Evaluating the effectiveness of retrieval-augmented large language models in scientific document reasoning. *arXiv preprint arXiv:2311.04348*.

Sai Munikoti, Anurag Acharya, Sridevi Wagle, and Sameera Horawalavithana. 2024. Atlantic: Structure-aware retrieval-augmented language model for interdisciplinary science. In *Proceedings of the Workshop on AI to Accelerate Science and Engineering (AI2ASE)*, Vancouver, Canada. Held in conjunction with the 38th AAAI Conference on Artificial Intelligence (AAAI 2024).

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.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Anthropic Team and Collaborators. Article about claude 3 models. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2024-05-06.

Gemini Team and Collaborators. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Sridevi Wagle, Sai Munikoti, Anurag Acharya, Sara Smith, and Sameera Horawalavithana. 2024. Empirical evaluation of uncertainty quantification in retrieval-augmented language models for science. In *Proceedings of the Workshop on Scientific Document Understanding (SDU)*, Vancouver, Canada. Held in conjunction with the 38th AAAI Conference on Artificial Intelligence (AAAI 2024).

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*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. 2024. $infty$ bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*.

A Question Definitions

NEPA experts reviewed and created the definitions for each question types as following.

1. **Closed questions:** Closed questions have two possible answers depending on how you phrase it: "yes" or "no" or "true" or "false."
2. **Comparison questions:** Comparison questions are higher-order questions that ask listeners to compare two things, such as objects, people, ideas, stories or theories.
3. **Convergent questions:** convergent questions are designed to try and help you find the solution to a problem, or a single response to a question.
4. **Divergent questions:** Divergent questions have no right or wrong answers but rather encourage open discussion. While they are similar to open questions, divergent questions differ in that they invite the listener to share an opinion, especially one that relates to future possibilities.
5. **Evaluation questions:** Evaluation questions, sometimes referred to as key evaluation questions or KEQs, are high-level questions that are used to guide an evaluation. Good evaluation questions will get to the heart of what it is you want to know about your program, policy or service.
6. **Funnel questions:** Funnel questions are always a series of questions. Their sequence mimics a funnel structure in that they start broadly with open questions, then segue to closed questions.
7. **Inference questions:** Inference questions require learners to use inductive or deductive reasoning to eliminate responses or critically assess a statement.
8. **Problem-solving questions.** Problem-solving questions present students with a scenario or problem and require them to develop a solution.
9. **Process questions:** A process question allows the speaker to evaluate the listener's knowledge in more detail.
10. **Recall questions:** A recall question asks the listener to recall a specific fact.

B Sample Questions

In this sections, we listed the sets of sample questions we used for each types of questions.

B.1 Closed questions

- Are there any federally recognized Tribes in a 50-mile radius of [PROJECT]?
- Are there any federally recognized species of concern in a 50-mile radius of [PROJECT]?
- Did [AGENCY] approve the licensing action
- Did the EIS consider [SUBJECT]?

B.2 Comparison questions

- Which Tribes were consulted in [PROJECT 1] and not [PROJECT 2] and vice-versa?
- What are some differences between [STUDY 1] and [STUDY 2] that might account for differences in species count for [SPECIES]?
- Compare the considered alternatives in [PROJECT 1] with those in [PROJECT 2].
- Compare the outcomes of surveys from the new reactor EIS with the license renewal EIS for [RPOJECT].

B.3 Convergent questions

- Which other species of concern could logically be in within the 50-mile radius around the [PROJECT]?
- How many similar projects could be built before the impact level for air quality was rated as high?
- If the area of effect for the proposed action were increased by 50%, what additional federal species of concern would need to be addressed?

B.4 Divergent questions

- What considerations should the [AGENCY] addressed in the document but didn't?

B.5 Evaluation questions

- Based on NEPA evaluations done in the vicinity of [PROJECT], does the conclusion of the Historical and Cultural resources section appropriately weigh the concerns of Tribal leaders?

- Extrapolating using this and other NEPA evaluations, what is the long term outlook for [SPECIES] in the vicinity?
- How have [AGENCY'S] NEPA reviews trended over time and would this review have the same outcome 10 years ago or 10 years from now?
- In the license renewal EIS for [PROJECT], which impacts have changed from the initial EIS and why?

B.6 Funnel questions

- Which federally recognized Tribes are in a 50-mile radius of [PROJECT]? Which Tribes participated in this EIS? What were the concerns fo participating Tribes? What mitigations were made?
- Which federally recognized species of concern are in a 50-mile radius of [PROJECT]? What mitigations, if any, were made to project those species?
- Which alternatives were discussed? Which were considered? Why was [ALTERNATIVE] not considered?
- Which resource areas were discussed in the Affected Environmnent section of the document?
- What were the impacts of the proposed action on [SUBJECT]?
- Did [AGENCY] consider [X] when evaluting [SUBJECT]?

B.7 Inference questions

- If the federally recognized [TRIBE] has land in the vicinity of [PROJECT 1] like it does in the vicity of [PROJECT 2], what concerns might [TRIBE] have with [PROJECT 1]?
- If the primary migitation for [SPECIES] for [PROJECT TYPE] in the past has been [MITIGATION], what would you expect the mitigation to be for [PROJECT]?
- If [AGENCY 1] and [AGENCY 2] typically agree on impact levels and [AGENCY 1] found large impact in terrestrial ecology for an action in a nearby area, what would [AGENCY 2] find?
- If mitigations for air quality for [PROJECT 1] were effective and the same mitigations were applied to [PROJECT 2], what would we assume the outcome to be for [PROJECT 2]?

B.8 Problem-solving questions

- Given the following references, evaluate the effect of a new nuclear plant at [SITE] on cultural and historic resources in the vicinity.
- Given the location of the [PROJECT], create a list of aquatic species likely present in a 50-mile radius.
- Write an Abstract for [PROJECT]
- Given the list of reference in [SECTION] of [PROJECT 1] create a list of references applicable to [PROJECT 2]. Provide hyperlinks and ML numbers, if available.

B.9 Process questions

- How does this document define the NEPA process for consultation with Tribes?
- How does [AGENCY] define the area of effect for the proposed action?

B.10 Recall questions

- What references did [AGENCY] use in evaluating the effect of the applicant's proposed action on [SPECIES]?
- Which resource areas indicated a moderate or large impact due to the proposed action?

C EIS Dataset

Table 3 reports the statistics of the EIS data that used to create the benchmark.

Document Title	Agency	#Pages	#Tokens
Continental United States Interceptor Site	Missile Defense Agency, Department of Defense	74	41,742
Supplement Analysis of the Final Tank Closure and Waste Management for the Hanford Site, Richland, Washington, Offsite Secondary Waste Treatment and Disposal	Hanford Site Office, Department of Energy	63	43,167
Nationwide Public Safety Broadband Network Final Programmatic Environmental Impact Statement for the Southern United States	Department of Commerce	86	43,985
T-7A Recapitalization at Columbus Air Force Base, Mississippi	United States Department of the Air Force (DAF), Air Education and Training Command (AETC).	472	179,697
Oil and Gas Decommissioning Activities on the Pacific Outer Continental Shelf	The Bureau of Safety and Environmental Enforcement (BSEE) and Bureau of Ocean Energy Management (BOEM)	404	271,545
Final Environmental Impact Statement for the Land Management Plan Tonto National Forest	Department of Agriculture, Forest Service	472	325,641
Final Environmental Impact Statement for Nevada Gold Mines LLC's Goldrush Mine Project, Lander and Eureka Counties, NV	Bureau of Land Management, Interior.	454	413,083
Addressing Heat and Electrical Upgrades at Fort Wainwright, Alaska	Department of the Army, Department of Defense	618	514,003
Sea Port Oil Terminal Deepwater Port Project	The U.S. Coast Guard (USCG) and Maritime Administration (MARAD), Department oF Transportation	890	613,214

Table 3: Statistics on the EIS documents used in the evaluation