Examining Long-Context Large Language Models for Environmental Review Document Comprehension

Anonymous ACL submission

Abstract

As LLMs become increasingly ubiquitous, researchers have tried various techniques to augment the knowledge provided to these models. Long context and retrieval-augmented generation (RAG) are two such methods that have recently gained popularity. In this work, we examine the benefits of both of these techniques by utilizing question answering (QA) task in a niche domain. While the effectiveness of LLM-based QA systems has already been established at an acceptable level in popular domains such as trivia and literature, it has not often been established in niche domains that traditionally require specialized expertise. We construct the NEPAQuAD1.0 benchmark to evaluate the performance of five long-context LLMs – Claude Sonnet, Gemini, GPT-4, Llama 3.1, and Mistral – 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. 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 models in handling different types of questions (e.g., problem-solving, divergent, etc.). Our results suggest that RAG powered models significantly outperform those provided with only the PDF context in terms of answer accuracy, regardless of the choice of the LLM. Our further analysis reveals that many models perform better answering closed type questions (Yes/No) than divergent and problem-solving questions.

1 Introduction

As Large Language Model (LLM)s 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 Protection Act (NEPA)[1]. NEPA is a U.S. environmental law 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 prepared 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 LLMs and Retrieval Augmented Generation (RAG) techniques to evaluate the extent to which LLMs can comprehend NEPA knowledge for question-answering (Figure 1). We select frontier LLMs for our experiments: Claude Sonnet (Anthropic Team and Collaborators, 2024), Gemini (Gemini Team and Collaborators, 2024), GPT-4 (OpenAI, 2024), Llama 3.1(Dubey et al., 2024) and Mistral(Jiang et al., 2023a). 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 method employing GPT-4. Overall, we make the following contributions:

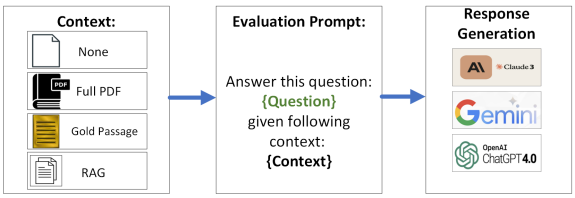

Figure 1: Illustration of varied EIS contexts used in the comparison.

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

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

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

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 two 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 type. 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

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

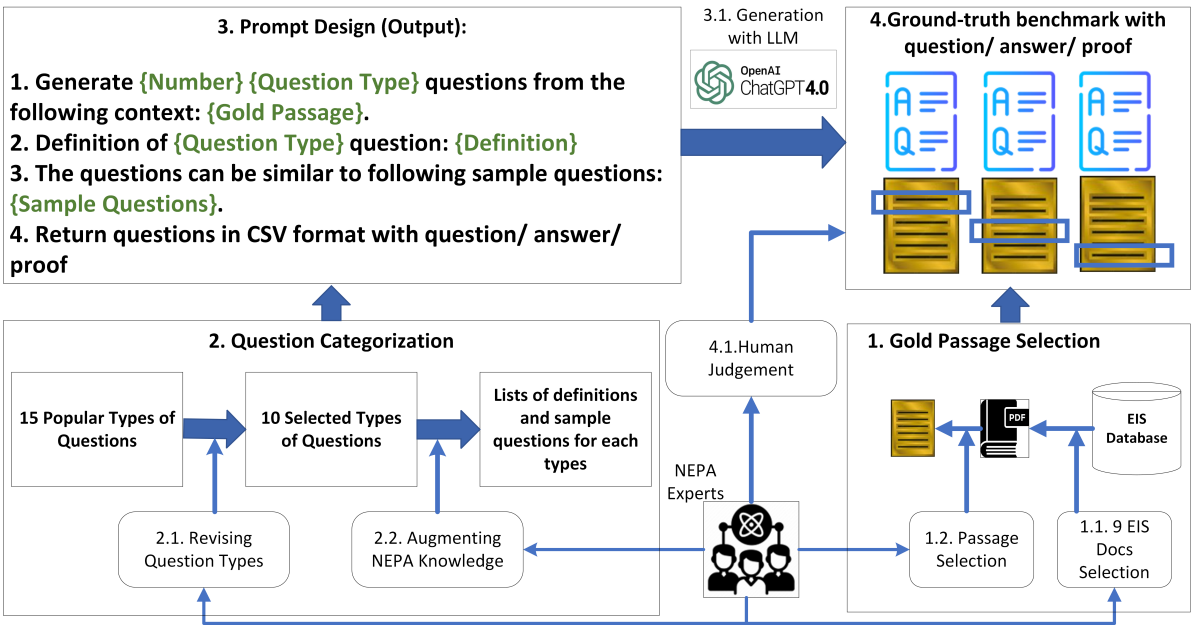

Figure 2: Steps of Ground-truth benchmark generation for evaluating LLMs over varied contexts for question-answering over EIS documents

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.

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 closed questions and two each of the other nine types of questions. Out of these total 2520 generated questions, a significant number of these were not outputted in the correct format by the LLMs. We then filtered the generations that were not in the desired formats, bringing the number down to 1920. Finally, we manually filtered the remaining questions for quality, thus ending with a benchmark of 1599 ground-truth triples of question-answer-proof over the nine selected EIS documents. We make the benchmark publicly available.[4]

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. While these human checks certainly helped assure the quality, we nevertheless still have some limitations in the dataset creation process, as discussed in Section 7.

3 Experimental Setup

Our experimental framework is designed to comprehensively evaluate the performance of various LLMs in responding to questions related to NEPA documents. We selected a diverse set of models, including both closed and open-source options, to provide a broad perspective on current LLM capabilities in this domain. The models under evaluation are: GPT-4 (OpenAI) (OpenAI, 2024), Claude-3 Sonnet (Anthropic) (Anthropic Team and Collaborators, 2024), Gemini (Google) (Gemini Team and Collaborators, 2024), 8B LLAMA3.1

[4] Link omitted for anonymity during review

(Meta)(Dubey et al., 2024) and 7B Mistral (Mistral AI) (Jiang et al., 2023a). To assess these models' performance across different scenarios, we developed a benchmark consisting of questions derived from NEPA documents. Each model will be evaluated under four distinct context conditions, allowing us to analyze how different levels of information availability affect their responses. The comprehensive evaluation framework will allow us to draw meaningful conclusions about the strengths and limitations of different LLMs in handling NEPA-related queries under varying levels of contextual information.

3.1 Context Variation

We designed four distinct context scenarios that simulate various real-world information retrieval situations. These scenarios are designed to test the models' capabilities under different levels of available context:

No Context: In this baseline scenario, questions are presented to the models without any additional context. While this strategy can work well with popular domains such as sport or literature, we assume that the NEPA domain may be challenging for the LLMs to provide accurate answers. In other words, this setting can be considered a test of the LLMs' ability to answer out-of-general-domain questions. As such, this setting is usually expected to return low performance.

PDF Context: In this scenario, in addition to the question, we also provide the model with text contents from the entire PDF document, from which the benchmark question was generated. Since we do not inform the model which part of the document to look at, the accuracy of generated responses 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 the 'no context' scenario. This setup evaluates the models' ability to process and extract relevant information from lengthy, complex documents.

Augmented Context: Employing a RAG approach, this setup provides the top-k most relevant text chunks retrieved from a vector database containing text chunks of NEPA documents. In our RAG model implementation, when a question is inputted for LLM generation, the corresponding context is extracted as a relevant passage from a given NEPA document. We use the standard RAG setup where we encode both question and retrieved passages with the BGE embedding model (Xiao et al., 2023). Cosine similarity scores are used to assess the similarity between the question and the contexts. In our case, the number of top-ranked relevant text chunks extracted is set at $k = 3$. This scenario assesses the models' performance when given highly relevant, focused information that is specifically retrieved to answer the question at hand.

Gold Context: In this configuration, we include the actual context from which the question was generated in the prompt, alongside the question content. This scenario represents an ideal case where the most relevant information is perfectly identified and provided. While the situation where users manually identify the correct text 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. This setup serves as an upper bound for performance, testing the models' ability to comprehend and utilize perfectly relevant information.

3.2 Evaluation Metrics

To assess the performance of the models across different configurations, we employ a sophisticated evaluation approach that goes beyond simple syntactic similarity measures. While overlap based metrics such as BLEU scores (Papineni et al., 2002) have been used in prior works, they are not ideal for evaluations where semantic accuracy is crucial. Therefore, we have chosen to use the RAGAS evaluation framework (Es et al., 2023), specifically focusing on the *answer correctness* score. The *answer correctness* score combines two key aspects – (i) *factual correctness*: this measures the accuracy of the generated answer at the phrase/clause level compared to the ground truth. We utilize GPT-4 to quantify the factual overlap between the generated answer and the expected answer (Es et al., 2023). (ii) *semantic correctness*: this is calculated by comparing the vector embeddings of the expected answers and the predicted answers using the BGE embedding model (Xiao et al., 2023). In our evaluation, we assign different weights to these aspects – 75% for *factual correctness* (weight of 0.75), and 25% for *semantic correctness* (weight of 0.25). This weighting allows us to prioritize factual accuracy while still considering the overall semantic similarity of the responses. By using this comprehensive evaluation metric, we aim to pro-

vide a more nuanced and accurate assessment of the LLMs' performance in handling NEPA-related queries across the different context scenarios. This approach enables us to capture both the precise factual content and the broader semantic accuracy of the generated responses, offering valuable insights into the models' capabilities in processing and utilizing relevant document information.

We employ a separate evaluation method for closed type questions which have only 'yes' or 'no' answers using the following steps. First, we process the responses generated by the LLMs to extract only the 'yes' or 'no' answer from the generated response. Any additional explanatory text or elaboration is removed from the response. Thereafter, the cleaned binary response is compared directly with the ground truth answer. If the cleaned response exactly matches the ground truth, it is assigned a full score of 100%. If there is any discrepancy, including cases where the model fails to provide a clear "Yes" or "No" answer, the response is scored as 0%. This straightforward evaluation method for closed questions allows us to assess the models' ability to provide accurate binary responses without being influenced by any uncertainties introduced by the RAGAS evaluation framework. We refer this metric as *answer correctness* score for closed type questions.

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 five long-context LLMs: GPT-4, Claude-3 Sonnet, Gemini, Mistral and Llama3.1 across various QA contexts (Section 4.1). Second, we compare the model performance across various question types (Section 4.2). 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 GPT-4 model produces the most accurate results by far. However, when PDF documents are provided as context, Claude surpassing GPT-4 and Gemini in the answer 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.

Table 2: Answer correctness of LLM responses on the NEPAQuAD benchmark across different context types. The best-performing setting for each model is shown in bold.

Context	None	PDF	RAG	Gold
GPT-4	67.00%	63.70%	74.36%	**76.60%**
Claude	64.53%	66.46%	75.16%	**76.84%**
Gemini	62.84%	65.90%	75.46%	**81.15%**
Mistral	64.95%	61.81%	72.88%	**75.34%**
Llama3.1	66.35%	59.52%	**74.01%**	72.73%

Open-source models (LLama3.1 and Mistral) generally underperform compared to closed LLMs across most context types, including PDF, RAG, and gold contexts. However, in the no-context scenario, all models exhibit similar performance, suggesting that the availability and quality of context significantly influence the performance gap between open-source and closed models.

Overall, RAG models perform better in comparison to the models provided with PDF documents as additional contexts. In RAG setup, the Gemini model outperforms other models in term of correctness, although the correctness scores of Claude and Gemini models are much closer. There is notable increase in Llama3.1's performance in the RAG setup when compared with the PDF contexts. This suggests that the performance of long-context models can be improved when provided with the most relevant context, as in the RAG setup.

As expected, all models perform best on average when provided with the gold passage in comparison to other context variations, as in this scenario, models 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. The only exception to this trend is the Llama3.1 model, which outperforms in the RAG setup as compared to the gold passage.

4.2 Evaluating Different Question Types

In this section, we summarize few interesting observations when analyzing the performance of the models across different question types. The 'process', 'comparison' and 'evaluation' questions generally benefit from richer contexts, with closed-source models demonstrating a more significant

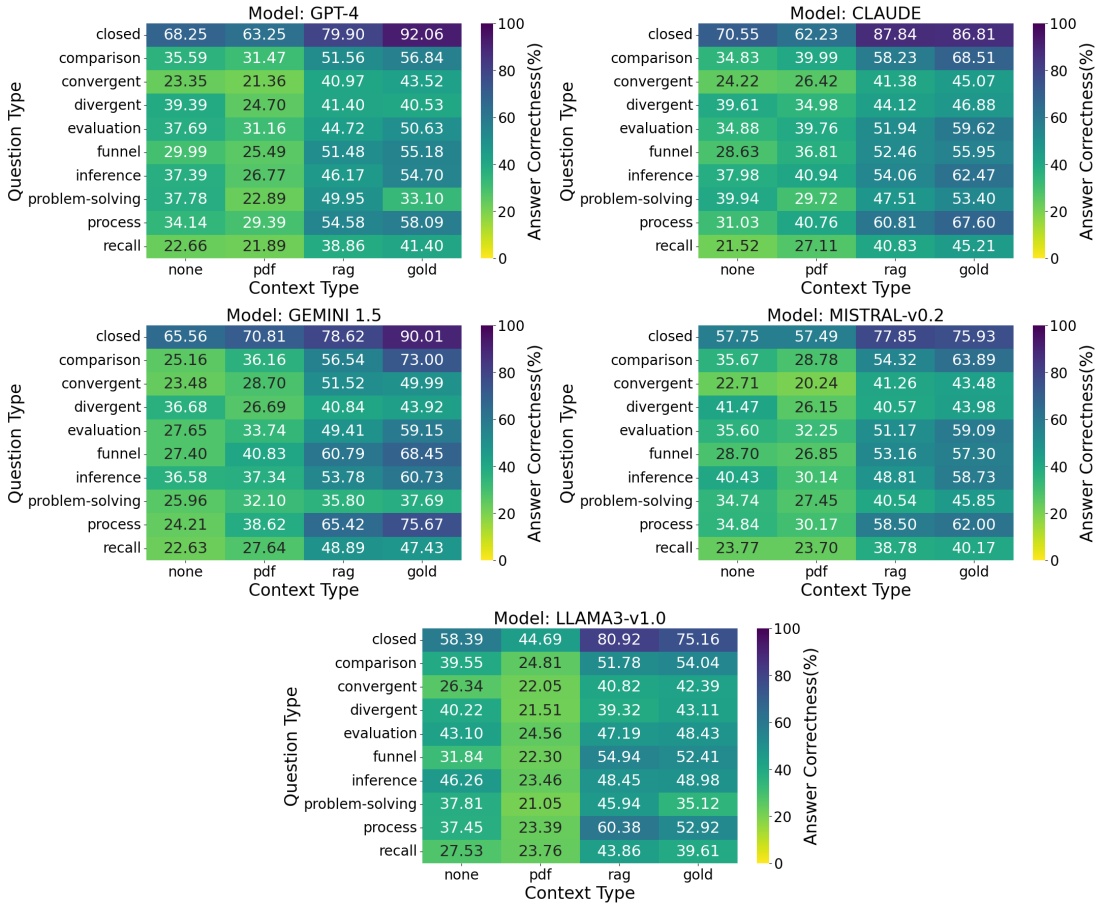

Figure 3: The evaluation results meassured by the Answer Correctness scores of each LLM used with 4 scenarios of using context over each question types

improvement. 'Convergent' and 'recall' type questions prove challenging for all models, especially in no-context scenarios, highlighting the difficulty of synthesizing information or retrieving specific details without supporting context. 'Problem-solving' questions show varied results across models, with some struggling even with rich context, indicating that complex analytical tasks remain a challenge. 'Funnel' questions benefit significantly from context, especially for models like Gemini 1.5, suggesting that structured, narrowing inquiries are well-suited to context-enhanced responses. The 'closed' questions yield the highest accuracy for all models, regardless of context. In this regard, it is important to note that the revised evaluation metric for closed questions, which directly compares the model's 'yes'/'no' response to the ground truth in contrast to the RAGAS metric used for other question types.

5 Related Work

Long Context Evaluation Recent works have shown that LLMs perform poorly in processing and understanding long, context-rich sequences in rigorous scientific benchmarks (Li et al., 2024; Liu et al., 2024). 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. Similarly, other studies showed that long-context LLMs are affected by the position of the relevant information in the input context. 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, while Liu et al. (2024) showed that LLMs perform weakly when they must access relevant information in the middle of long contexts. Another major work in evaluating LLM's long context understanding is Bai et al. (2023), which 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.

Likewise, L-Eval Benchmark (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. The ∞Bench work (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., 2023b) 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 NEPA domain 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 five long-context LLMs (both closed- and open-source) designed for handling extensive contexts 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. However, when these models are augmented with additional context, as seen in the RAG setup, their performance can be significantly improved. We encourage the research community to focus on advancing the complementary use of long-context LLMs and RAG techniques together, rather than treating them as distinct, separate approaches.

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 times 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.

Random errors when using RAGAS We noticed that RAGAS evaluation framework (Es et al., 2023) produces random errors and "NaN" values in specific types of question. While running experiments to analyze the details of this fall outside of scope of this work, our preliminary tests lead us to assume that it might be due to the way the framework interacts with generative models and tries to conform their output to its formats.

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*.

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

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: A case study in wind energy siting ordinances. *Energy and AI*, page 100431.

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*.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

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

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.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023a. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. 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.

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.

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