Benchmarking LLMs for Environmental Review and Permitting

Anonymous Author(s)

ABSTRACT

The National Environment Policy Act (NEPA) stands as a foundational piece of environmental legislation in the United States, requiring federal agencies to consider the environmental impacts of their proposed actions. The primary mechanism for achieving this is through the preparation of Environmental Assessments (EAs) and, for significant impacts, comprehensive Environmental Impact Statement (EIS). Large Language Model (LLM)s' effectiveness in specialized domains like NEPA remains untested for adoption in federal decision making processes. To address this gap, we present **NEPA Qu**estion and **A**nswering **D**ataset (*NEPAQuAD*), the first comprehensive benchmark derived from EIS documents, along with a modular and transparent evaluation pipeline *MAPLE* to assess LLM performance on NEPA focused regulatory reasoning tasks. Our benchmark leverages actual EIS documents to create diverse question types, ranging from factual to complex problem-solving ones. We built a modular and transparent evaluation pipeline to test both closed- and open-source models in zero-shot or context-driven QA benchmarks. We evaluate five state-of-the-art LLMs – Claude Sonnet 3.5, Gemini 1.5 Pro, GPT-4, Llama 3.1, and Mistral-7B-Instruct – using our framework to assess both their prior knowledge and their ability to process NEPA-specific information. The experimental results reveal that all the models consistently achieve their highest performance when provided with the gold passage as context. While comparing the other context-driven approaches for each model, Retrieval Augmented Generation (RAG)-based approaches substantially outperform PDF document contexts indicating that neither model is well suited for long-context question-answering tasks. Our analysis suggests that NEPA focused regulatory reasoning tasks pose a significant challenge for LLMs, particularly in terms of understanding the complex semantics and effectively processing the lengthy regulatory documents.

KEYWORDS

large language models, retrieval augmented generation, benchmarks, environmental permitting

ACM Reference Format:
Anonymous Author(s). 2025. Benchmarking LLMs for Environmental Review and Permitting. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25).* ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '25, August 03–07, 2025, Toronto, CA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

LLMs are demonstrating increasingly sophisticated cognitive abilities, including idea generation [9], problem-solving [35], and the potential to simulate believable human behaviors such as reflection and planning [32]. These capabilities are enabling LLMs to assist significantly in professional workflows and data-driven decision-making across various sectors, notably in Law [36] and Medicine [24]. For example, LLMs are already assisting with tasks such as drafting documents, reviewing information, and generally speeding up rote, repetitive, or data-intensive tasks within the decision making processes that are prone to human error or consume significant time [15, 28]. However, despite this demonstrated potential and application in other professional domains, their widespread adoption in government decision-making processes remains limited, primarily due to concerns surrounding trust and reliability.

In this work, we assess LLM performance in the domain of environmental reviews conducted under the National Environment Policy Act (NEPA[1]). NEPA stands as a foundational piece of environmental legislation in the United States, requiring federal agencies to consider the environmental impacts of their proposed actions. Beyond simply protecting the environment, the NEPA process plays a important role in promoting societal progress and responsible economic growth. By assuring informed decision-making, promoting transparency, and incorporating public input, NEPA helps shape federal decisions that are more sustainable, resilient, and responsive to community needs. This proactive approach can prevent costly environmental impacts, facilitate the development of necessary infrastructure in an environmentally sound manner, and build public trust, ultimately contributing to long-term economic stability and community well-being.

However, the NEPA decision making process is inherently complex, and characterized by different agency-wide regulations, multidisciplinary technical analysis, and a deeply collaborative process involving diverse stakeholders including applicants, consultants, regulatory agencies, legal teams, and the public[2]. To perform such decisions, it requires more than just comprehensive knowledge of federal, state, and local laws; it demands the ability to interpret their application to specific site conditions, analyze complex environmental interactions, predict potential outcomes, and communicate findings clearly and effectively. NEPA decision makers want the LLMs to perform the higher-order regulatory focused reasoning beyond fact retrieval for effective decision support. To support the adoption of LLMs in NEPA decision making workflows[3], we test the LLM's abilities to perform regulatory focused reasoning to interpret the intent behind regulations, apply their principles effectively, and support logical deductions or evaluations grounded in NEPA regulatory frameworks. This involves evaluating potential compliance pathways, identifying likely environmental impacts

[1] https://www.epa.gov/nepa
[2] https://www.epa.gov/nepa/national-environmental-policy-act-review-process
[3] https://www.whitehouse.gov/presidential-actions/2025/04/updating-permitting-technology-for-the-21st-century/

based on regulatory criteria, recommending appropriate mitigation strategies, and predicting plausible agency review outcomes based on project characteristics.

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[4]. 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.

To this end, we evaluate the extent to which LLMs can comprehend NEPA knowledge for diverse question typologies specifically designed to assess NEPA focused regulatory reasoning over EIS. We select frontier LLMs for our experiments: Claude Sonnet [3], Gemini-1.5Pro [13], GPT-4 [30], Llama 3.1[8] and Mistral-7B-Instruct[21]. We built **NEPA Qu**estion and **A**nswering **D**ataset (*NEPAQuAD*) benchmark that includes the triplets of questions, answers, and corresponding contexts, generated through a hybrid approach employing GPT-4 and NEPA experts. To support our evaluation, we developed a **M**ulti-context **A**ssessment **P**ipeline for **L**anguage model **E**valuation (MAPLE), a transparent and modular evaluation pipeline that seamlessly handles response generation across different LLM providers. MAPLE also supports various evaluation scenarios – basic question-answering with no context, and advanced features like RAG-based response generation and context-grounded evaluation. Using the MAPLE framework, we conduct rigorous experiments evaluating frontier LLMs with various types of contexts for NEPA documents. Overall, we make the following contributions:

(1) Constructed the first-ever benchmark NEPAQuAD v1.0 to evaluate the performance of LLMs for diverse question typologies specifically designed to assess NEPA focused regulatory reasoning.
(2) Developed MAPLE as a standardized evaluation pipeline to compare the capabilities of LLMs in various context-driven prompting strategies (i.e., gold passage, entire PDF document, and RAG) to assess model performance.
(3) NEPA experts driven performance evaluation in the LLMs ability to support real-world regulatory decisions.

The structure of this paper is outlined as follows. In Section 2, we describe the NEPAQuAD v1.0 benchmark in detail. Section 3 lays out our approach and the various contexts used for evaluation implementation, followed by a detailed analysis of our performance in Section 4. Section 5 then discusses other socio-scientific benchmarks and evaluation frameworks in the literature. We finish with the conclusion and limitations of our work in Sections 6 and 7.

2 THE NEPAQUAD V1.0 BENCHMARK

We present NEPAQuAD v1.0 (**N**ational **E**nvironmental **P**olicy **A**ct **Qu**estion and **A**nswering **D**ataset Version 1.0), the first-ever benchmark to evaluate the performance of LLMs in a question-answering task for EIS documents. NEPAQuAD consists of 1590 questions, split into two different question types: open and closed. The open questions are further split into nine different types, as shown in Table 1. Overall, these questions not only require LLMs to process and comprehend long-context documents, but also reason in the field of environmental permitting. As such, these questions are perfectly suited to serve as the test for any LLM's capability to reason in this regulatory domain.

2.1 Dataset Creation

Due to the high costs associated with manually creating the entire dataset and the inability to use ground-truth benchmarks from other domains, we adopt a hybrid approach with GPT-4 and NEPA experts to generate the NEPAQuAD v1.0 benchmark. The process we followed for this benchmark generation is illustrated in Figure 1, which we describe in more detail in the following sections:

Document Curation The first task is to curate a diverse set of high-quality EIS documents from federal databases and select excerpts to establish a representative evaluation corpus. NEPA experts selected nine EIS documents from different government agencies that were most representative of various NEPA actions. These documents exhibit significant variation 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 U.S. Army Corps of Engineers could prioritize water resource development and infrastructure impacts. Please refer to Table 6 in the Appendix for a brief overview about the selected documents. We consider documents of varied lengths, with the longest document containing nearly 900 pages (>600K tokens).

Gold Passage Selection For each of the nine selected documents, we attempt to select excerpts that have important content of each document. A naive approach of randomly extracting passages poses the risk of resulting in *low-quality* excerpts, such as parts of appendices or image 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, resulting a total of 10 excerpts from each document. We use these excerpts as the ground-truth context, called *gold passages*, for question benchmark generation.

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

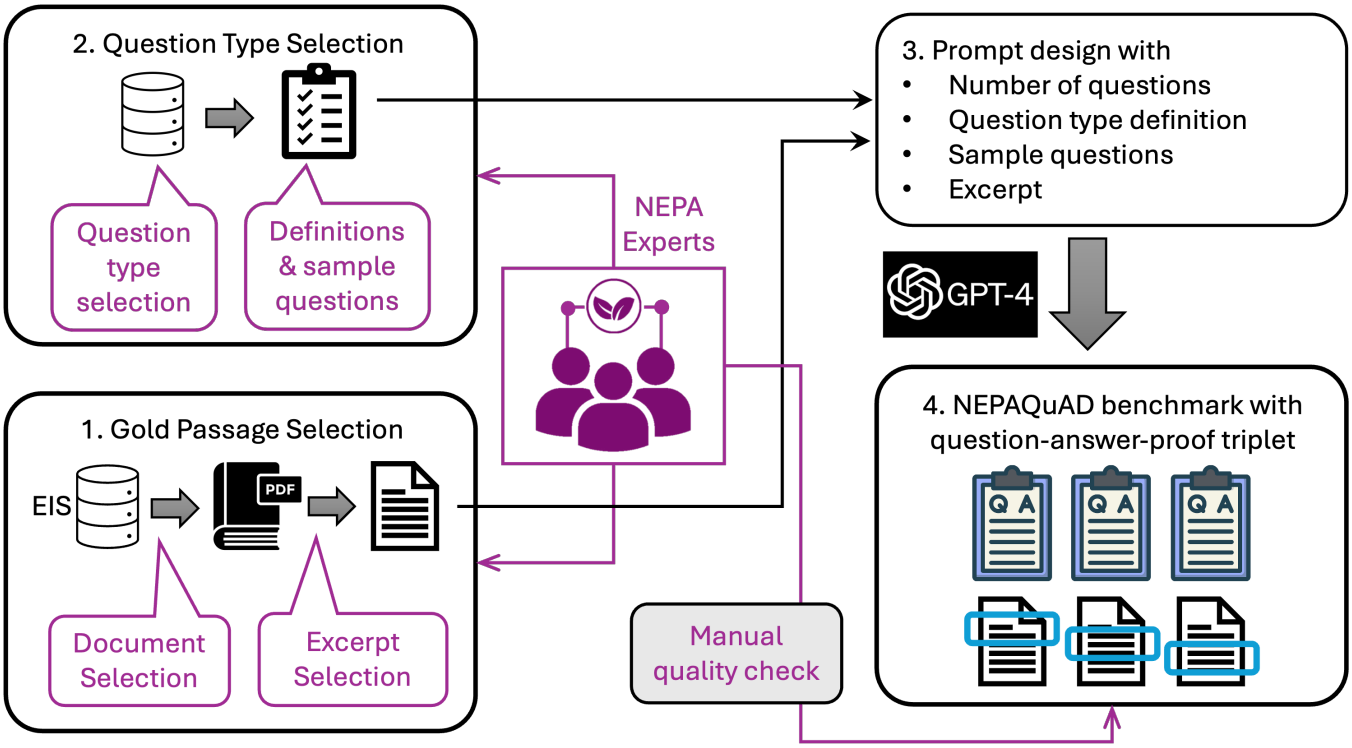

Figure 1: Steps of Ground-truth benchmark generation for evaluating LLMs over varied contexts for question-answering over EIS documents

Question Type Selection Next, we developed a question taxonomy that encompasses various complexities inherent in environmental impact assessments. After extensive discussions with NEPA experts, we finalized 10 types of questions that facilitate the NEPA decision making process. These question types test the LLM's ability to articulate the purpose of regulations, explain complex review processes, or describe the steps involved in a permit application.

For example, *funnel* questions regarding the environmental alternative analysis, such as *'Which alternatives were discussed?', 'Which were considered?' 'Why was [ALTERNATIVE] not considered?'* require in-depth environmental and legal knowledge to justify these decisions. On the other hand, problem solving questions such as *'Given the location of the [PROJECT], create a list of aquatic species likely present in a 50-mile radius'* present specific regulatory scenarios to test the ability of the models to comprehend the latest geographical knowledge. Overall, we ensure almost half the questions in the dataset are these types of 'open' questions. We present the question types and relevant examples 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 provided by the NEPA experts, please refer to Appendices B and C.

Prompt Design We leverage generative models (such as GPT-4) with carefully crafted prompts to generate question-answer-proof triplets based on the selected EIS contexts, ensuring coverage across different question types. To ensure that the prompts can instruct generative models efficiently, we consulted with 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 provided in Appendix A.

Table 1: Question types in the NEPAQuAD v1.0 benchmark

Type	#Questions		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%

Automated Generation Following successful precedents in other domain specific benchmark creation [4], we employed GPT-4 through Azure OpenAI service with default parameters to generate the question-answer-proof benchmark triplets. Our dataset construction began with nine carefully selected EIS documents, from which we extracted 10 gold passages each, totaling 90 passages. For each passage, we generated a diverse set of questions: 10 closed-type questions and two questions each from nine other question categories. While the initial generation yielded 2520 questions, we implemented a rigorous quality control process. First, we filtered out responses that did not conform to our specified output format, reducing the count to 1920. Subsequently, we conducted manual quality assessment of the remaining questions, resulting in a final

Context: Definition of Resource Socioeconomics pertains to the social and economic conditions of the human environment. For this analysis, the indicators of socioeconomic conditions include population, employment, unemployment rate, income, cost of living, and housing availability. Current data on these indicators for the affected environment would provide the baseline information on the socioeconomic well-being of the local areas/region upon which potential effects of the proposed project alternatives are compared. The proposed heat and electrical generation and distribution facility upgrades would occur within Fort Wainwright, located in the FNSB. The borough is the cultural and commercial center of the Interior Region as well as a hub for villages located hundreds of miles outside the region (Alaska Department of Labor and Workforce Development [ADOLWD] 2018)

Closed

Question: Does the definition of resource include biological studies?
Answer: No.

Comparison

Question: How do the roles of Fort Wainwright and the FNSB differ with respect to their regions in Alaska?
Answer: Fort Wainwright is a military facility undergoing upgrades within the FNSB, which is the cultural and commercial center of the Interior Region and acts as a hub for remote villages. Fort Wainwright is located within the FNSB but caters specifically to military functions, while the FNSB serves a wider role in the region.

Divergent

Question: What future cultural or commercial impacts might arise from the heat and electrical generation and distribution facility upgrades at Fort Wainwright, considering its status as a regional hub?
Answer: Potential impacts could range from attracting new businesses and boosting the local economy to affecting the cultural demographics and social dynamics within Fort Wainwright and the surrounding villages, given its role as the cultural and commercial center of the Interior Region.

Process

Question: What process is outlined for establishing the baseline information on the socioeconomic well-being of the local areas/region affected by the proposed project?
Answer: The process for establishing the baseline information involves collecting current data on indicators such as population, employment, unemployment rate, income, cost of living, and housing availability for the affected environment. This baseline is used for comparing the potential effects of the proposed project alternatives.

Inference

Question: If Fort Wainwright's proposed project serves as a cultural and commercial hub for outlying villages, how might the facility upgrades affect these external areas?
Answer: The facility upgrades could potentially affect the socioeconomic well-being of the villages that relay on Fort Wainwright as a hub, due to interconnected economic and cultural activities.

Funnel

Question: What is the role of the FNSB as mentioned in the document?
Answer: The FNSB is the cultural and commercial center of the Interior Region and a hub for villages located hundreds of miles outside the region.

Figure 2: Examples of question-answer pairs present in NEPAQuAD v1.0 , and the context used to generate these questions.

benchmark of 1590 high quality question-answer-proof triplets across the selected EIS documents. Few examples of question-answer pairs generated for a gold passage context is illustrated in Figure 2. NEPAQuAD v1.0 is made publicly available [33] to facilitate further research and development in this domain.

2.2 Quality Assessment

Comparison with other QA datasets. We test the intrinsic quality of the generated questions with several metrics related to content and style, and compare the metrics against the similar QA datasets SQuAD [34] and SciQ [39]. The following metrics are used to assess data quality:

(1) Entailment between the context and the question and answer, measured via prediction from a BERT model fine-tuned on natural language inference (NLI)[5];
(2) Complexity of the question, measured via parse tree depth from a spacy dependency parser (`en_core_web_trf` model);
(3) Specificity of the question, measured via number of named entities extracted using a SpaCy[6] named entity recognition model.
(4) Readability of the question, measured via Flesch-Kincaid readability score [25].

The results in Table 2 show that NEPAQuAD v1.0 is on-par or exceeding in the metrics as compared to similar QA datasets. For instance, the questions in NEPAQuAD v1.0 are more complex than SQuAD and SciQ, while the proportion of questions with entailment or neutral NLI labels is in a similar range to the other datasets. The lower readability score for NEPAQuAD v1.0 indicates longer sentences and a more complex vocabulary, which is reasonable considering the NEPA domain.

Table 2: Comparing NEPAQuAD with existing QA benchmarks using various question quality metrics

Metric	NEPAQuAD	SQuAD	SciQ
% Entailment + Neutral	0.925	0.985	0.968
Complexity	5.317	3.867	4.741
Specificity	1.17	1.11	0.139
Readability	34.4	60.5	58.0

Human Review To ensure a high-quality benchmark, we employed a rigorous annotation protocol. We randomly sampled 100 benchmark entries from the entire benchmark dataset. Two independent NEPA experts evaluated each sampled entry across three critical dimensions: the correctness of question type (i.e. whether the generated question was the same type of question as requested), the correctness of the generated answer, and the correctness of generated proof. For each aspect, annotators provided binary judgments ('yes'/'no') based on established criteria documented in the annotation guidelines provided to the experts. For entries where the

[5]Accessed 1 April 2025: https://huggingface.co/tasksource/ModernBERT-base-nli
[6]https://spacy.io/

Table 3: Summary of annotation agreement among three annotators.

Aspect	% Majority Yes	% Majority No	% Unanimous Agreement
Is question type correct?	81.0	19.0	89.0
Is the answer correct?	93.0	7.0	71.0
Is the proof correct?	93.0	7.0	79.0

Table 4: Annotator agreement scores with ground truth (majority rule)

		Is question type correct?	Is the answer correct?	Is the proof correct?
NEPA expert 1	% Agreement	99.0	81.0	87.0
	Cohen's Kappa	0.967 (Almost Perfect)	0.328 (Fair)	0.340 (Fair)
NEPA expert 2	% Agreement	90.0	90.0	92.0
	Cohen's Kappa	0.734 (Substantial)	0.423 (Moderate)	0.560 (Moderate)

NEPA experts disagreed on one or more aspects, we engaged a third independent NEPA expert to adjudicate, creating a three-annotator subset. To determine overall quality and establish ground truth, we then employed a majority voting approach. These scores are shown in Table 3. Furthermore, we evaluated each annotator's alignment with this majority opinion by calculating individual Cohen's Kappa scores between each expert's responses and the majority vote as shown in Table 4. We show examples of human annotation of the generated triplet in Appendix D.

3 THE MAPLE EVALUATION FRAMEWORK

We developed MAPLE (Multi-context Assessment Pipeline for Language model Evaluation) as a comprehensive evaluation tool that streamlines the assessment of LLMs in question answering and document retrieval tasks. As illustrated in Figure 3, we designed it to provide a unified interface supporting multiple LLM providers while ensuring consistent evaluation methodologies. We incorporated a modular design that we used to integrate various context types – from no-context baselines to RAG-enhanced setups. We also implemented RAGAS evaluation metrics [10], allowing us to benchmark LLM performance across different benchmarks. The MAPLE architecture comprises four modules: a flexible *dataloader* that standardizes benchmark entries from various file formats, a provider-agnostic *LLM handler* enabling unified interaction across multiple LLM services, an *evaluator* module that dynamically populates prompt templates with questions and context information from benchmark entries, and a comprehensive *metrics* module supporting both user-defined custom metrics and established metric evaluation frameworks like RAGAS [2]. We discuss the key features of MAPLE below. Please refer to Appendix G for a detailed overview of individual modules.

3.1 Supporting multiple LLM providers

We implemented an *LLM Handler* module using an abstract base class to provide support for major cloud services. Our implementation includes Azure OpenAI (GPT models), AWS Bedrock (Claude, LLama, Mistral), Google Vertex AI (Gemini), and local HuggingFace models. We designed the handler to manage provider-specific authentication requirements and parameter specifications while maintaining a standardized interface for response generation. In this way, we effectively abstract the complexities of working with different LLM services into a single, consistent evaluation pipeline.

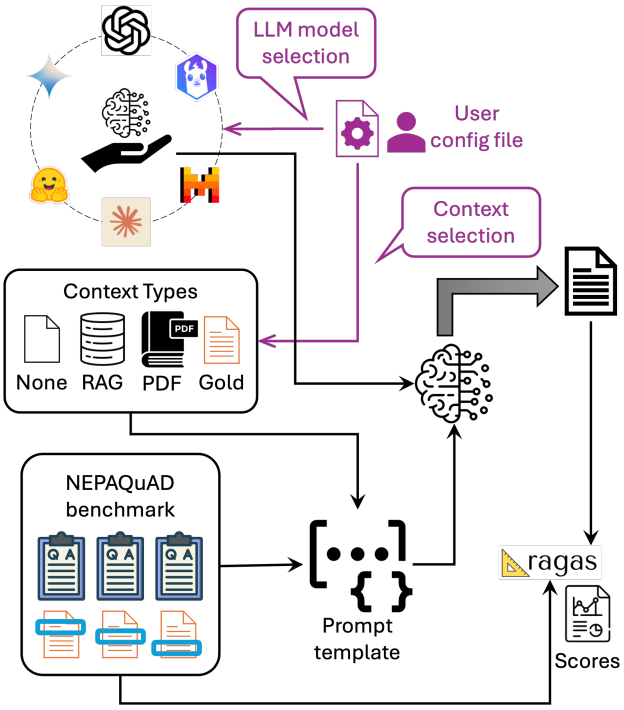

Figure 3: Overview of the MAPLE pipeline.

3.2 Supporting multiple context types.

The *Evaluator* module (which generates LLM responses to the questions in the benchmark) supports four context types to enable comprehensive evaluation across different scenarios. These include a basic *no-context* mode to test prior knowledge of LLM, a *PDF* mode for processing complete documents, a *RAG* [27] mode for vector database integration, and a *gold* context mode for controlled evaluations with the gold passage as context. MAPLE automatically identifies and validates appropriate context types by analyzing the available fields in benchmark entries. It defaults to the 'no-context'

mode when only question-answer pairs are present. Here, we provide brief description of these four context types.

No Context: Questions are presented to the models without any additional context. As such, one can think of this mode as analogous to using an LLM-based chatbot. 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, for our benchmark, 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 addition to the question, we also provide the model the text from the entire PDF document from which the 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.

RAG 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 [40]. 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 text excerpt 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.

3.3 Supporting custom prompt templates

We designed the prompt engineering system of MAPLE to be flexible and customizable through a template-based approach. We implemented a system where users can provide prompt templates through external files, with designated placeholders for *question* and *context* elements. The system automatically formats these templates during evaluation by replacing placeholders with actual content. We also included validation checks to ensure required placeholders are present for specific context types. For example, we require both question and context placeholders for *gold* context mode evaluations, while basic *no-context* evaluation requires only the placeholder for question. An example prompt template with placeholders for question and context is shown below. Through this design, we enabled users to experiment with different prompting strategies while ensuring consistent prompt structures across evaluations.

Prompt template with placeholder

You are a helpful assistant who will answer questions about environmental policies from a context provided in the prompt.
Question: {placeholder for question}
Answer the question from the context.
Context: {placeholder for context}

3.4 Evaluation metrics

We leveraged the RAGAS evaluation framework [10] to assess LLM responses against benchmark ground truth answers. We specifically focused on the *answer correctness* score, which is configured to combine two key components: factual correctness and semantic correctness. For factual correctness, which comprises 75% of the final score, we utilized GPT-4 to evaluate the accuracy of generated answers at the phrase/clause level compared to ground truth answer in the benchmark. We complemented this with semantic correctness, weighted at 25%, where we employed the BGE embedding model [40] to compare vector embeddings of predicted and expected answers. We chose this comprehensive evaluation approach over traditional overlap-based metrics like BLEU scores[31], as it efficiently captures both factual accuracy and semantic similarity of the generated responses and ground truth answers.

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 examine the performance of the LLMs on NEPAQuAD v1.0 using the MAPLE framework (as presented in Section 3). First, we compare the performance of the five frontier LLMsacross various contexts (Section 4.1). Then, we compare the model performance across various question types (Section 4.2).

4.1 Evaluating Different QA Contexts

Table 5 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.

Table 5: Answer correctness of LLM responses on the NEPAQuAD v1.0 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%

Despite the fact 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. 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

Figure 4 shows 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 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. It is important to note that the evaluation metric for closed questions 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 WORKS

Benchmarking LLMs for socio-scientific domains In recent years we have witnessed the emergence of numerous domain-specific benchmarks designed to rigorously evaluate LLMs across specialized fields. In the context of environment and environmental sciences, notable contributions include benchmark suite for water engineering [41], the ELLE dataset [17], the EnviroExam benchmark [20], and the WeQA benchmark [29] focusing on RAG-based evaluation in the wind energy domain. Furthermore, specialized benchmarks for professional domains have also surfaced such as the LawBench benchmark [5], CFinBench [18], and TransportationGames [44] to evaluate knowledge of LLMs in legal, financial and transport domains respectively. SciEval [19] is another example which provides a multi-level evaluation benchmark for scientific research capabilities of LLMs. SciBench [38] and FrontierMath [14] are premiere benchmarks to assess scientific and mathematical problem-solving capabilities of LLMs. More recently, PhysReason [43], QASA [26] and the CURIE dataset [6] have become popular benchmark dataset to evaluate LLM performance on scientific reasoning tasks.

Frameworks for LLM Evaluations The evaluation of LLMs has recently evolved beyond simple task-based metrics toward more sophisticated frameworks that address specific evaluation challenges across diverse domains. More recently, complementary approaches to tackle different aspects of LLM evaluation have emerged [22, 42]. Efforts to standardize evaluation processes have produced several notable toolkits, including Evalverse [23], FMEval [11], and LLMeBench [7].For specialized RAG-based frameworks, Guinet et al. [16] presents an automated evaluation protocol utilizing task-specific exam generation. One of the most famous evaluation frameworks is LM Eval Harness [12], which provides a unified platform for measuring LLM performance across hundreds of standardized tasks and has become a cornerstone for reproducible model evaluation in the research community. Similarly, RAGAS [2] provides multiple types of metrics to evaluate LLM responses on dimensions like faithfulness and context relevance, but does not provide a comprehensive workflow framework for end-to-end evaluation. DeepEval [1] offers testing capabilities for LLM applications but is primarily optimized for OpenAI models, potentially limiting its applicability across different model providers. Our proposed MAPLE framework offers a transparent and modular approach that seamlessly integrates multiple LLM providers, supports customizable prompt templates, handles various context types, and incorporates diverse evaluation metrics within a unified evaluation pipeline.

6 DISCUSSION AND 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

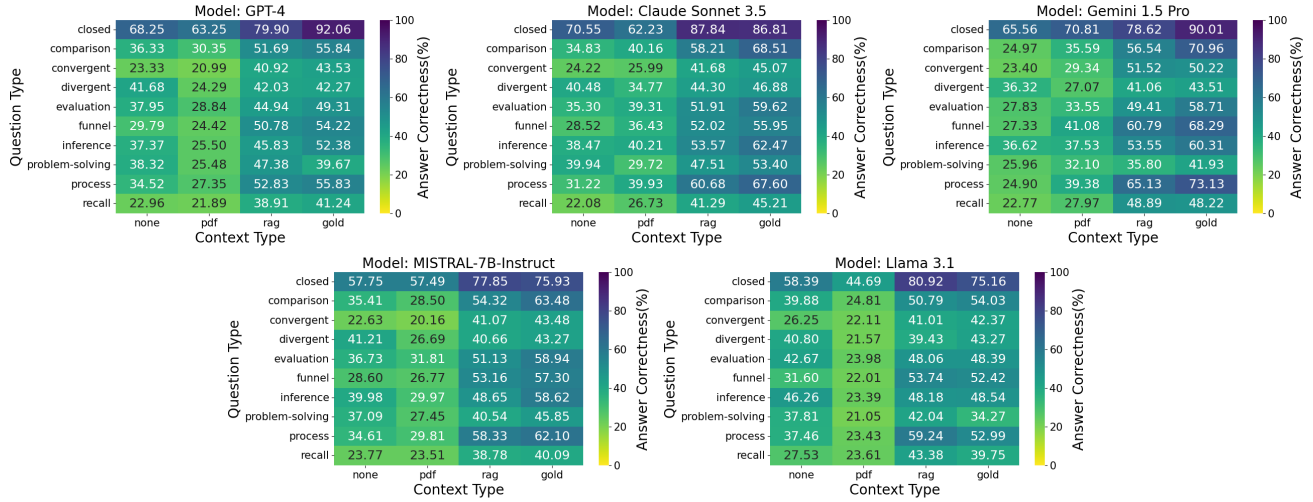

Figure 4: The evaluation results meassured by the Answer Correctness scores of each LLM used with 4 scenarios of using context over each question types

closed- and open-source) designed for handling extensive contexts across various contextual settings.

Our analysis indicates that NEPA regulatory reasoning tasks 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 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 models 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.

7 LIMITATIONS

As with all work, 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 needed to truncate the content of Full PDF to run these two models. This might cause the performance drop on the full PDF context mode. In future work, we should analyze more carefully about the impact of token truncation.

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, as LLMs might generate different responses each time, even with the same input, as demonstrated in another study [37]. Multiple runs of the same model might mitigate this effect.

Challenges of human judgment. Currently, we leverage human evaluation only on a subset of entries in the benchmark. With sufficient time and resources, a thorough evaluation of the entire dataset following the same methodology will likely make the dataset even more robust.

Imbalance of question types. In the current benchmark, just around half the questions are closed type. This is due to a myriad of reasons, including difficulty in generating open questions and the time-consuming nature of open question evaluation. While open-questions do take significantly more human review and effort, a future benchmark with even more such questions will undoubtedly provide an even tougher challenge to LLMs.

Random errors when using RAGAS We noticed that RAGAS evaluation framework [10] 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 [40] 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.

REFERENCES

[1] 2023. DeepEval – The LLM Evaluation Framework. https://github.com/confident-ai/deepeval.

[2] 2023. RAGAS. https://docs.ragas.io/en/stable/.

[3] 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.

[4] 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. arXiv:2401.08406 [cs.CL]

[5] Kai Chen, D. Zhu, Jidong Ge, Zhiwei Fei, Zhuo Han, Xiaoyu Shen, Zongwen Shen, Fengzhe Zhou, and Songyang Zhang. 2023. LawBench: Benchmarking Legal Knowledge of Large Language Models. *ArXiv* abs/2309.16289 (2023). https://api.semanticscholar.org/CorpusId:263134950

[6] Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, et al. 2025. CURIE: Evaluating LLMs On Multitask Scientific Long Context Understanding and Reasoning. *arXiv preprint arXiv:2503.13517* (2025).

[7] Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur A. Chowdhury, Hamdy Mubarak, Ahmed M. Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2023. LLMeBench: A Flexible Framework for Accelerating LLMs Benchmarking. In *Conference of the European Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusId:260735890

[8] 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* (2024).

[9] Daniel Nygård Ege, Henrik H Øvrebø, Vegar Stubberud, Martin Francis Berg, Christer Elverum, Martin Steinert, and Håvard Vestad. 2024. ChatGPT as an inventor: Eliciting the strengths and weaknesses of current large language models against humans in engineering design. *arXiv preprint arXiv:2404.18479* (2024).

[10] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217 [cs.CL]

[11] Luca Franceschi, Muhammad Bilal Zafar, Pinal Tailor, Pola Schwöbel, Michael Diamond, Michele Donini, Keerthan Vasist, Tomer Shenhar, Pinar Yilmaz, and Aman Malhotra. 2024. Evaluating Large Language Models with fmeval. *ArXiv* abs/2407.12872 (2024). https://api.semanticscholar.org/CorpusId:271270071

[12] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The Language Model Evaluation Harness. doi:10.5281/zenodo.12608602

[13] Gemini Team and Collaborators. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL]

[14] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872* (2024).

[15] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems* 36 (2023), 44123–44279.

[16] Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. *ArXiv* abs/2405.13622 (2024). https://api.semanticscholar.org/CorpusId:269983057

[17] Jing Guo, Nan Li, and Ming Xu. 2025. Environmental large language model Evaluation (ELLE) dataset: A Benchmark for Evaluating Generative AI applications in Eco-environment Domain. *ArXiv* abs/2501.06277 (2025). https://api.semanticscholar.org/CorpusId:275471371

[18] Tianyu Guo, Weijian Sun, Ying Nie, Wei He, Binfan Zheng, Qiang Li, Hao Liu, Binwei Yan, Dacheng Tao, Yunhe Wang, Weihao Wang, and Haoyu Wang. 2024. CFinBench: A Comprehensive Chinese Financial Benchmark for Large Language Models. *ArXiv* abs/2407.02301 (2024). https://api.semanticscholar.org/CorpusId:270878688

[19] Yang Han, Baocai Chen, Da Ma, Lu Chen, Liangtai Sun, Zihan Zhao, Kai Yu, and Zhe-Wei Shen. 2023. SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific Research. *ArXiv* abs/2308.13149 (2023). https://api.semanticscholar.org/CorpusId:261214653

[20] Yu Huang, Liang Guo, Wanqian Guo, Zhe Tao, Yang Lv, Zhihao Sun, and Dongfang Zhao. 2024. EnviroExam: Benchmarking Environmental Science Knowledge of Large Language Models. *ArXiv* abs/2405.11265 (2024). https://api.semanticscholar.org/CorpusId:269921863

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

[22] Shafiq Joty, Yixin Liu, Alexander Fabbri, Peifeng Wang, Chien-Sheng Wu, Yilun Zhao, Arman Cohan, and Kejian Shi. 2024. ReIFE: Re-evaluating Instruction-Following Evaluation. *ArXiv* abs/2410.07069 (2024). https://api.semanticscholar.org/CorpusId:273228598

[23] Jihoo Kim, Wonho Song, Dahyun Kim, Yunsu Kim, Yungi Kim, and Chanjun Park. 2024. Evalverse: Unified and Accessible Library for Large Language Model Evaluation. *ArXiv* abs/2404.00943 (2024). https://api.semanticscholar.org/CorpusId:268819292

[24] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems* 37 (2024), 79410–79452.

[25] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (1975).

[26] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*. PMLR, 19036–19052.

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[28] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. *arXiv preprint arXiv:2403.16812* (2024).

[29] Rounak Meyur, Hung Phan, S. Wagle, Jan Strube, M. Halappanavar, Sameera Horawalavithana, Anurag Acharya, and Sai Munikoti. 2024. WeQA: A Benchmark for Retrieval Augmented Generation in Wind Energy Domain. In *unknown*. https://api.semanticscholar.org/CorpusId:271915856

[30] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[31] 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*. 311–318.

[32] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[33] PolicyAI. 2024. LLM for Environmental Review. https://kaggle.com/competitions/llm-for-environmental-review

[34] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

[35] Sumedh Rasal and EJ Hauer. 2024. Optimal decision making through scenario simulations using large language models. *arXiv preprint arXiv:2407.06486* (2024).

[36] Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access* (2025).

[37] 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)..

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

[39] Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 94–106.

[40] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[41] Yue Yang, Boyan Xu, How yong Ng, Xiongpeng Tang, Rui Tong, Zihao Li, Xueqing Shi, Liang Wen, Yu Li, Yuxin Yang, Qingxian Su, Zihao Wu, and Guanlan Wu. 2024. Unlocking the Potential: Benchmarking Large Language Models in Water Engineering and Research. *ArXiv* abs/2407.21045 (2024). https://api.semanticscholar.org/CorpusId:271571554

[42] Wenjin Yao, Jindong Wang, Zhuohao Yu, Chang Gao, Yidong Wang, Shikun Zhang, Xing Xie, Wei Ye, and Yue Zhang. 2024. KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/

CorpusId:267897557
[43] Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. 2025. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054* (2025).
[44] Xue Zhang, Xiangyu Shi, Xinyue Lou, Rui Qi, Yufeng Chen, Jinan Xu, and Wenjuan Han. 2024. TransportationGames: Benchmarking Transportation Knowledge of (Multimodal) Large Language Models. *ArXiv* abs/2401.04471 (2024). https://api.semanticscholar.org/CorpusId:266900046

A PROMPT TEMPLATE

The template for the prompt is shown below:

Prompt template with placeholder

1. Generate {placeholder for number of questions} {placeholder for type of question} questions from the following context.
Context: {placeholder for gold passage}
2. Definition of {placeholder for type of question} question: {placeholder for definition provided by NEPA experts}
3. The questions can be similar to following sample questions: placeholder for sample questions provided by NEPA experts
4. Return the questions in CSV format with columns question / answer / proof.

We provided appropriate placeholders for the type and number of questions to be generated, the excerpt from which the question will be generated along with the sample questions of the same type as provided by the NEPA experts. 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.

B 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.

C SAMPLE QUESTIONS

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

C.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]?

C.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].

C.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?

C.4 Divergent questions

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

C.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?

C.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]?

C.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]?

C.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.

C.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?

C.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?

D HUMAN ANNOTATION PROCESS

The first example shows a triplet which was accepted by both NEPA experts (shown in green), whereas the second example shows an example where there was a conflict of opinions between the NEPA experts (shown in orange).

NEPA expert evaluation of GPT-4 generated triplet

Type: process
Question: How does the document define the socioeconomic conditions for the analysis of the human environment?
Answer: The document defines socioeconomic conditions for the analysis as indicators including population, employment, unemployment rate, income, cost of living, and housing availability.
Proof: Socioeconomics pertains to the social and economic conditions of the human environment. For this analysis, the indicators of socioeconomic conditions include population, employment, unemployment rate, income, cost of living, and housing availability.
NEPA expert 1 evaluation: Correct triplet.
NEPA expert 2 evaluation: The triplet is correct.
Overall decision: Include in NEPAQuAD v1.0

NEPA expert evaluation of GPT-4 generated triplet

Type: Divergent
Question: What future cultural or commercial impacts might arise from the heat and electrical generation and distribution facility upgrades at Fort Wainwright, considering its status as a regional hub?
Answer: Potential impacts could range from attracting new businesses and boosting the local economy to affecting the cultural demographics and social dynamics within Fort Wainwright and the surrounding villages, given its role as the cultural and commercial center of the Interior Region.
Proof: The borough is the cultural and commercial center of the Interior Region as well as a hub for villages located hundreds of miles outside the region.
NEPA expert 1 evaluation: The type of generated question is correct. The answer and proof are incorrect. Table ES-1 in the document provides a summary of socioeconomic impacts that would improve the answer. The answer also conflates Fort Wainwright as the regional hub, whereas the actual text of the EIS identifies the FNSB as the cultural hub.
NEPA expert 2 evaluation: Provided triplet is correct.
Overall decision: Consult with a third NEPA expert.

E EIS DATASET

Table 6 reports the statistics of the EIS data that used to create the benchmark.

Table 6: Statistics on the EIS documents used in the evaluation

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

F EXAMPLES FROM NEPAQUAD

Here we present an example of a gold passage which was selected by NEPA experts from an EIS document on the Fort Wainwright generation and distribution facility.

Selected Excerpt

3.5.1.1 Definition of Resource Socioeconomics pertains to the social and economic conditions of the human environment. For this analysis, the indicators of socioeconomic conditions include population, employment, unemployment rate, income, cost of living, and housing availability. Current data on these indicators for the affected environment would provide the baseline information on the socioeconomic well-being of the local areas/region upon which potential effects of the proposed project alternatives are compared. The proposed heat and electrical generation and distribution facility upgrades would occur within Fort Wainwright, located in the FNSB. The borough is the cultural and commercial center of the Interior Region as well as a hub for villages located hundreds of miles outside the region (Alaska Department of Labor and Workforce Development [ADOLWD] 2018).

Now we present some examples of questions from NEPAQuAD which have been generated using our methodology in Section 2 and accepted to be included by the NEPA experts. These examples include the question-answer-proof triplets which are generated using our methodology.

- **Closed** question: 'Does the definition of resource include biological studies?', answer: 'No', proof: 'Socioeconomics pertains to the social and economic conditions of the human environment.'
- **Comparison** question: 'How do the roles of Fort Wainwright and the FNSB differ with respect to their regions in Alaska?', answer: 'Fort Wainwright is a military facility undergoing upgrades within the FNSB, which is the cultural and commercial center of the Interior Region and acts as a hub for remote villages. Fort Wainwright is located within the FNSB but caters specifically to military functions, while the FNSB serves a wider role in the region.', proof: 'The proposed heat and electrical generation and distribution facility upgrades would occur within Fort Wainwright, located in the FNSB. The borough is the cultural and commercial center of the Interior Region as well as a hub for villages located hundreds of miles outside the region.'
- **Divergent** question: 'What future cultural or commercial impacts might arise from the heat and electrical generation and distribution facility upgrades at Fort Wainwright, considering its status as a regional hub?', answer: 'Potential impacts could range from attracting new businesses and boosting the local economy to affecting the cultural demographics and social dynamics within Fort Wainwright and the surrounding villages, given its role as the cultural and commercial center of the Interior Region.', proof: 'The borough is the cultural and commercial center of the Interior Region as well as a hub for villages located hundreds of miles outside the region.'
- **Funnel** question: 'What is the role of the FNSB as mentioned in the document?', answer: 'The FNSB is the cultural and commercial center of the Interior Region and a hub for villages located hundreds of miles outside the region.', proof: 'The borough is the cultural and commercial center

of the Interior Region as well as a hub for villages located hundreds of miles outside the region.'

- **Inference** question: 'If Fort Wainwright's proposed project serves as a cultural and commercial hub for outlying villages, how might the facility upgrades affect these external areas?', answer: 'The facility upgrades could potentially affect the socioeconomic well-being of the villages that relay on Fort Wainwright as a hub, due to interconnected economic and cultural activities.', proof: 'The borough is the cultural and commercial center of the Interior Region as well as a hub for villages located hundreds of miles outside the region.'
- **Process** question: 'What process is outlined for establishing the baseline information on the socioeconomic well-being of the local areas/region affected by the proposed project?', answer: 'The process for establishing the baseline information involves collecting current data on indicators such as population, employment, unemployment rate, income, cost of living, and housing availability for the affected environment. This baseline is used for comparing the potential effects of the proposed project alternatives.', proof: 'Current data on these indicators for the affected environment would provide the baseline information on the socioeconomic well-being of the local areas/region upon which potential effects of the proposed project alternatives are compared.'

G MAPLE MODULES

G.1 LLM Handlers

The LLM Handler infrastructure provides a unified interface for interacting with various Language Model providers. This module implements an abstract base class that standardizes the interaction with different LLM services while accommodating provider-specific requirements. The key features of the LLM Handler module are discussed below.

Supported providers. The LLM handler of MAPLE supports the major cloud providers:

- Azure OpenAI services with GPT models
- AWS Bedrock with Claude, LLama and Mistral models
- Google Vertex AI with Gemini models
- Local deployment of HuggingFace models

Authentication management. The authentication management system in MAPLE's LLM Handlers provides provider-specific authentication while maintaining consistent security standards. Azure OpenAI services utilize API key validation and endpoint configuration, while AWS Bedrock supports both role-based credentials through AWS STS and profile-based authentication through AWS SSO. Google Vertex AI implements service account authentication using JSON key files with automatic credential refresh mechanisms. For local HuggingFace models, there is no authentication system, rather it only requires the model path. The system implements credential caching, automatic retries for authentication failures, and secure credential storage across all handlers, effectively abstracting provider-specific authentication complexities from the evaluation pipeline.

Response generation capabilities. The LLM Handlers provide a standardized interface for generating responses across different providers while maintaining provider-specific optimizations. All handlers implement consistent response formatting, error handling for failed generations, and proper resource cleanup, ensuring reliable response generation regardless of the underlying provider.

G.2 Utility Modules

The Utility Modules provide essential support functions for data management, logging, and visualization. These modules form the foundation for data processing and result analysis within the framework.

Data Loading and Processing. The data loading system provides structured handling of benchmark datasets and evaluation results. A key feature is its automatic context type detection based on CSV column availability: presence of the 'file_name' column enables both RAG and PDF context types as it allows document linkage, while the 'context' column enables Gold context type by providing ground truth contexts. The basic 'none' context type is always supported as it requires only question and answer columns. The system enforces data validation for required fields and handles optional attributes flexibly. It processes both input benchmark questions and evaluation results, maintaining data consistency and providing clear feedback on data quality issues through a hierarchical exception handling system. This context type detection allows downstream components to automatically determine valid evaluation strategies based on the available data.

Logging Infrastructure. The logging infrastructure implements a centralized logging system through a singleton pattern, ensuring consistent log management across all components. It provides configurable log levels, automatic file rotation, and structured log formatting. The system maintains separate logs for different components while enabling unified log access, with automatic creation of log directories and files, and proper handling of concurrent logging requests.

Configuration Management. The configuration system handles various framework settings through a combination of YAML files and environment variables. It manages provider credentials, folder paths, and evaluation parameters, implementing validation checks and providing default values where appropriate. The system supports different configurations for various evaluation scenarios while maintaining security for sensitive information.

Type Validation. The type validation system enforces data consistency through strongly-typed data structures. It implements comprehensive validation for both required and optional fields, handles missing values appropriately, and provides clear error messages for validation failures. The system ensures data integrity across the framework while maintaining flexibility for different data formats.

G.3 Evaluator Module

The Evaluator Module manages the response generation process, implementing different context types and handling the evaluation pipeline. This module coordinates the interaction between LLM handlers and input data.

Context Type Management. The Evaluator module supports multiple context types for response generation, each with specialized processing pipelines. The supported context types are discussed here.

- `none` context type: The basic mode generates responses without additional context.
- `pdf` context type: The 'pdf' context mode processes entire documents, managing token limits and content extraction. The text chunks are loaded from the JSON files stored within the data directory (required input for this context type). In case of tokens exceeding the token limit, we define methods to truncate the text chunks to stay within the allowed limits.
- `rag` context type: This mode integrates with a vector database for relevant document retrieval and filtering. The path to the vector database and the collection name needs to be provided as inputs for this context type.
- `gold` context type: This mode utilizes predefined ground truth contexts for controlled evaluation scenarios.

The system automatically validates context availability and requirements for each mode, ensuring appropriate context handling based on available data.

Prompt Template Management. The Evaluator module supports customizable prompt engineering through a template-based system. Users can provide prompt templates with placeholders for 'question' and 'context' (when applicable) through a template file and provide its path as an input. The system automatically formats the prompt by replacing these placeholders with actual content during evaluation, raising appropriate warnings if required placeholders are missing for the chosen context type. For instance, templates for 'gold' context type must contain both question and context placeholders, while basic evaluation might only require question. This flexibility in prompt design allows users to experiment with different prompting strategies and maintain consistent prompt structures across evaluations without modifying the core pipeline.

Progress Management and Persistence. The evaluator implements a robust progress tracking and persistence system that processes benchmark questions sequentially while maintaining state. For each benchmark question, it generates a response, processes the result, and immediately stores it in the output CSV file. This immediate persistence strategy ensures that progress is never lost due to interruptions. When restarting an interrupted evaluation, the system automatically detects previously processed questions by checking the existing output file and continues from the last successful evaluation. This approach provides resilience against system failures and allows for long-running evaluations to be safely paused and resumed, while maintaining detailed logs of each processing step.

Response Management. The response management system handles the collection, validation, and storage of generated responses with their associated metadata. For RAG-based evaluations, it tracks source documents and relevance scores, while for context-based generations, it maintains context-response relationships. The system implements structured storage of responses, supporting both incremental updates and batch commits, with comprehensive logging of generation parameters and outcomes.

Error Recovery. The error recovery system provides robust handling of various failure scenarios during evaluation. It implements automatic retries for transient failures, and structured logging of unrecoverable errors. The system maintains evaluation progress despite individual question failures and provides detailed error reporting for post-evaluation analysis.

G.4 Metrics Module

The Metrics Module implements RAGAS evaluation metrics for assessing the quality of LLM responses. This module provides comprehensive evaluation capabilities across multiple dimensions of response quality.

Supported Metrics. The Metrics module implements a subset of RAGAS evaluation metrics for assessing LLM response quality. The current version supports five core metrics: answer correctness for measuring accuracy, answer similarity for response alignment, context precision and recall for evaluating context relevance, and answer faithfulness for assessing response consistency with provided contexts. Each metric provides a score between 0 and 1, enabling quantitative assessment of different response aspects.

Batch Evaluation. The module processes multiple evaluation responses in configurable batch sizes to optimize resource usage and evaluation speed. For each batch, it computes all selected metrics simultaneously using the RAGAS framework, managing memory efficiently by processing fixed-size subsets of the full evaluation dataset. This batched approach provides a balance between processing speed and resource consumption.

Progress Management. Similar to the Evaluator module, the Metrics module implements immediate result persistence by writing metric scores to the output CSV file after each batch evaluation. The system tracks progress through the evaluation set and supports continuation from the last successful evaluation in case of interruptions. This ensures reliable processing of large evaluation datasets while maintaining evaluation state.

NaN Recovery and Recomputation. The module provides functionality to identify and recompute metrics for responses where RAGAS evaluation resulted in NaN (Not a Number) values. It scans the results CSV file for NaN entries in specified metric columns, extracts the corresponding responses and contexts, and performs targeted re-evaluation of these specific entries. The recomputed values are then seamlessly integrated back into the original results file, maintaining data consistency while improving evaluation completeness.

Result Storage and Visualization. Evaluation results are stored in structured CSV files with consistent column ordering: file metadata columns followed by question details, expected and predicted responses, and individual metric scores. This standardized format enables direct integration with the visualization utilities, supporting generation of comparative heatmaps and performance bar plots. The results can be analyzed across different models, context types, and question categories using the visualization tools available within the utilities module of MAPLE.

Received 18 May 2025