SCORE: Specificity, Context Utilization, Robustness, and Relevance for Reference-Free LLM Evaluation

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used to support question answering and decision-making in high-stakes, domain-specific settings such as natural hazard response and infrastructure planning, where effective answers must convey fine-grained, decision-critical details. However, existing evaluation frameworks for retrieval-augmented generation (RAG) and open-ended question answering primarily rely on surface-level similarity, factual consistency, or semantic relevance, and often fail to assess whether responses provide the specific information required for domain-sensitive decisions. To address this gap, we propose a multi-dimensional, reference-free evaluation framework that assesses LLM outputs along four complementary dimensions: specificity, context utilization, robustness to paraphrasing and semantic perturbations, and answer relevance. We introduce ClimateRiskQA, a curated dataset of 1,412 domain-specific question–answer pairs spanning 40 professional roles and seven natural hazard types, and further evaluate on PubMedQA to examine cross-domain generalization. We also conduct human evaluation to assess inter-annotator agreement and alignment between model outputs and human judgments, which highlights the inherent subjectivity of open-ended, domain-specific evaluation. Our results show that no single metric sufficiently captures answer quality in isolation and demonstrate the need for structured, multi-metric evaluation frameworks when deploying LLMs in high-stakes applications.

1 Introduction

LLMs have shown remarkable capabilities across a wide range of tasks, especially in answer generation and decision support (Brown et al., 2020; Ouyang et al., 2022; Wei et al., 2022). As these models are increasingly deployed in high-stakes domains such as healthcare, legal reasoning, and climate risk assessment (Wu et al., 2025; Nagar et al., 2024; Li et al., 2024a; Luo et al., 2025; Li et al., 2024b); the need for robust and reliable evaluation methods has become essential. One such critical domain is natural hazard response, where LLMs are used to recommend mitigation strategies and provide factual information about disasters in specific geographic regions (Xie et al., 2025). In such scenarios, even minor inaccuracies can lead to real-world consequences.

Despite this growing reliance on LLMs, evaluating their outputs in the above-mentioned settings remains challenging. Traditional reference-based evaluation metrics are often insufficient, as they primarily measure surface-level similarity to gold-standard answers rather than verifying whether generated responses contain correct and specific information (Celikyilmaz et al., 2020; Kryściński et al., 2020; Chang et al., 2024). This limitation is particularly acute in natural hazard and extreme weather response, where publicly available gold-standard datasets are scarce due to the diversity of hazards, geographic regions, infrastructure systems, and professional contexts. Effective answers in this domain must be geographically precise, and tailored to specific concerns.

Recent work has shifted toward reference-free evaluation, including LLM-as-a-judge pipelines, and RAG evaluation frameworks. These methods evaluate general conversational quality, factuality, answer relevance, claim-level verification, and context relevance or utilization (Bai et al., 2024; Li et al., 2023; Chiang et al., 2024; Es et al., 2024; Ru et al., 2024; Saad-Falcon et al., 2024; Ni et al., 2024). Although effective for general settings, existing RAG evaluation metrics fall short in domain-specific settings. Faithfulness and factuality metrics assess whether answers are broadly supported by retrieved documents, but do not verify whether critical specific details are present and correct (Min et al., 2023). Answer relevance metrics mostly rely

on semantic similarity or LLM-based scores, which can overestimate quality in cases where responses are generic rather than informative and relevant (Liu et al., 2023). Claim-level verification focuses on factual correctness at the level of individual claims, but does not evaluate whether the answer sufficiently addresses the user's query intent or decision context. Similarly, prior context-utilization methods measure overlap between retrieved knowledge and generated answers, but do not fully determine whether the model actually relied on that information during generation (Ru et al., 2024).

To address these limitations, we propose SCORE, a multi-perspective evaluation framework for assessing LLM-generated answers in natural hazard analysis and decision support. SCORE evaluates responses along four complementary dimensions: *specificity*, *context utilization*, *robustness*, and *relevance*. For specificity, we employ a multi-evaluator LLM-as-a-judge setup to verify whether key details (e.g., hazard type, location, timeline, intensity) are explicitly stated and supported by retrieved evidence. For context utilization, we measure whether and how retrieved passages contribute to the generated answer. Our metric measures evidence utilization at a finer granularity through counterfactual dependence by removing individual claims and quantifying the resulting change in answer confidence, enabling us to distinguish necessary evidence from merely consistent, ignored, or distracting context. We assess robustness by applying paraphrasing and controlled perturbations to the question and measuring whether model responses remain consistent or change appropriately. For answer relevance, we adopt a question-regeneration strategy with an additional semantic masking step to reduce superficial keyword overlap and better capture conceptual alignment.

In this work, we present a reference-free evaluation framework SCORE for assessing LLM outputs in high-stakes hazard response. Our key contributions are as follows: (1) We construct ClimateRiskQA, a synthetic dataset of 1,412 domain-specific question–answer pairs that cover different hazards, locations, infrastructure types, and concerns to support controlled and scalable evaluation; (2) We design a multi-dimensional, reference-free evaluation framework that assesses specificity, context utilization, robustness, and answer relevance, and validate these metrics through human evaluation; and (3) We demonstrate the generalizability of SCORE beyond hazard response through evaluation on a clinically relevant subset of PubMedQA.

Code & Data: Our code and curated data are available at https://anonymous.4open.science/r/score-8D5F/readme.md

2 Dataset Construction

We construct ClimateRiskQA, a curated dataset of infrastructure-related questions that simulate real-world hazard scenarios and decision-making contexts. Recent work (Xie et al., 2025) demonstrates the effectiveness of role-aware RAG for wildfire risk mitigation. We extend this idea to a broader evaluation benchmark covering multiple infrastructure sectors, hazard types, and geographic settings.

User Profile. Each question–answer pair is associated with a synthetic user profile consisting of profession, concern type, location, and timeline. The concern type is fact-based, recommendation-seeking, or hybrid. Profile attributes are generated through controlled randomization: we first sample a hazard type, then a location conditioned on the hazard, followed by profession and timeline.

Professional Context. To ground questions in real-world professional contexts, we curate a diverse set of 40 professions spanning five critical infrastructure sectors: *transportation*, *water*, *energy*, *buildings*, and *communications* (Appendix Table 7). Each corresponds to roles that are actively involved in hazard mitigation and infrastructure planning, such as transportation planner, hydraulic engineer, energy storage specialist, building systems manager, and telecommunications engineer.

Location and Hazard Type. We identify a set of seven hazard types: *cold wave*, *heat wave*, *coastal flooding*, *ice storm*, *hurricane*, *drought*, and *wildfire*. Each hazard is then associated with a curated list of U.S. locations that are historically or geographically susceptible to it, based on the Natural Risk Index (Appendix Table 8).[1] This helps ensure that the generated questions reflect regionally relevant risks and real-world challenges faced by infrastructure planners.

Question Generation. Questions are generated from templates populated with infrastructure, hazard, location, concern, and timeline attributes. The concern type determines the template pool: fact-based questions target objective vulnerabilities, recommendation-seeking questions request mitigation strategies, and hybrid questions focus on interdependencies or cascading impacts.

[1] https://hazards.fema.gov/nri/data-resources

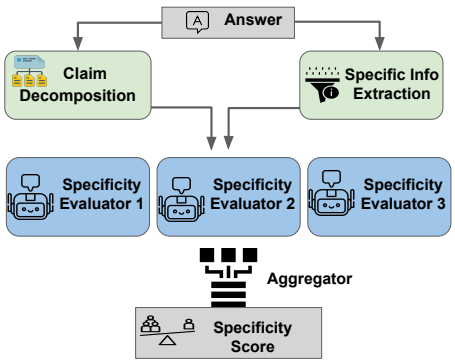


Figure 1: Specificity score computation: each generated answer is decomposed into atomic claims, specific details (hazard type, location, timeline, intensity) are extracted, and each claim is evaluated using multiple LLM-based evaluators before aggregating their judgments.

Answer Generation. For each question, we retrieve the top five semantically relevant abstracts from a domain-specific knowledge base (Mallick et al., 2023, 2025). These abstracts, along with the user profile and query, are incorporated into a structured prompt that is passed to an LLM to generate a concise, point-wise answer grounded in the retrieved literature (Appendix H.1).

3 Evaluation Framework

A major contribution of our work is the design of a reference-free, multi-dimensional evaluation framework for assessing LLM-generated answers in high-stakes, domain-specific settings. To rigorously evaluate the responses generated in Section 2, our framework measures the answer quality across four fine-grained dimensions: *specificity*, *context utilization*, *robustness*, and *answer relevance*. Additionally, we assess *readability* using established metrics to ensure that the outputs are suitable for professional use (Appendix C.4).

3.1 Specificity

We propose a multi-evaluator evaluation framework, illustrated in Figure 1, to measure the specificity of LLM-generated answers by assessing whether fine-grained details (e.g., hazard type, location, timeline, and intensity) are explicitly stated and verifiable against the retrieved knowledge. Formally, given the set of atomic claims $C := \{c_1, c_2, \ldots, c_{|C|}\}$ extracted from generated answers, and specificity dimensions $D := \{d_h, d_l, d_t, d_i\}$ representing **h**azard type, **l**ocation, **t**imeline, and **i**ntensity. Each evaluator k assigns a label $\ell_{id_j}^{(k)} \in \{\text{yes, no, n/a}\}$ to each claim c_i, along each dimension. Here, yes indicates that the specific detail is both present in the claim and verifiable using the knowledge source; no indicates a lack of support or contradiction; and n/a denotes that the detail is not mentioned in the claim. We map these categorical labels to numerical values:

$$s_{id_j}^{(k)} = \begin{cases} 1, & \text{if } \ell_{id_j}^{(k)} = \text{“yes”}, \\ 0, & \text{if } \ell_{id_j}^{(k)} = \text{“no”}, \\ \text{n/a}, & \text{if } \ell_{id_j}^{(k)} = \text{“n/a”}. \end{cases}$$

For each claim c_i and dimension d_j, we perform majority voting across k evaluators, to obtain:

$$\mathbf{s}_i = \left[s_{id_h},\ s_{id_l},\ s_{id_t},\ s_{id_i} \right].$$

For each dimension d_j, we average the consensus s_{id_j} across claims where $d_j \neq$ n/a. This prevents infrequently extracted dimensions (e.g., intensity) from being penalized by averaging over all claims:

$$\bar{s}_{d_j} = \frac{\sum_{i \in C\,:\, s_{id_j} \neq \text{n/a}} s_{id_j}}{\sum_{i \in C\,:\, s_{id_j} \neq \text{n/a}} 1}$$

The final specificity score is a weighted combination of dimension-level averages:

$$\text{Specificity}(C) = \frac{\sum_{j:\, \bar{s}_{d_j}} \alpha_j\, \bar{s}_{d_j}}{\sum_{j:\, \bar{s}_{d_j}} \alpha_j} \qquad (1)$$

where $\alpha = [0.6,\ 0.2,\ 0.1,\ 0.1]$ corresponds to the ordered dimensions: *hazard*, *location*, *timeline*, and *intensity*. If a dimension is “n/a” across claims, it is excluded from both the numerator and the denominator in the weighted average. In our domain, the four specificity dimensions do not contribute equally to assessing the quality of a claim. We place higher emphasis on the hazard and location dimensions, which are most critical for grounding claims in real-world risk contexts.[2]

3.2 Context-Utilization

We measure context utilization as the degree to which each unique claim in the retrieved context contributes to the model's final answer. Given a query q and an answer a, we first extract a set of unique claims $C := \{c_1, c_2, ...c_{|c|}\}$ from the retrieved documents D. To quantify the contribution

[2] Appendix C.1 provides additional details on claim and specific information extraction, and evaluator prompts.

of each claim $c_i \in C$ we adopt a leave-one-out uncertainty–based estimator that measures the change in answer confidence when c_i is removed from the context. Formally, we define the absolute contribution of claim c_i as the drop in model's confidence in generating the answer a:

$$\Delta_i(a, q, C, \theta) = f_\theta(a|q \cup C) - f_\theta(a|q \cup C \setminus \{c_i\})$$

where $f_\theta(a|.)$ denotes the confidence score of language model in generating a conditioned on q, C. We further define relative context utilization as:

$$\delta_i(a, q, C, \theta) = \Delta_i(a, q, C, \theta)/f_\theta(a|q \cup C)$$

Confidence Estimation. We compute answer confidence using the normalized per-token probability of the generated sequence (Murray and Chiang, 2018). Specifically, confidence is defined as the geometric mean of token probabilities:

$$f_\theta(a|X) = e^{\frac{1}{T}\sum_{t=1}^{T} log(P(y_t|X, y_{<t}))} \quad (2)$$

Where $a := (y_1, y_2, ...y_T)$ denotes the answer token sequence, and X denote the conditioning context ($q \cup C$). This formulation avoids length bias and numerical instability associated with raw probability multiplication.[3] Using this, $\Delta_i \in (-1, 1)$, with larger positive values indicating greater reliance on claim c_i. $\delta_i \in (-\infty, 1)$, where positive values indicate that the claim contributes positively to the answer, while negative values suggest that the claim is distracting or detrimental.

Finally, we aggregate claim-level contributions to obtain a global context utilization score that summarizes how effectively the retrieved context is used by the model when producing the answer. It is defined as the average (CU) or relative (CU_{rel}) contribution across all claims:

$$CU = \frac{1}{|C|}\sum_{i \in |C|} \Delta_i; \quad CU_{rel} = \frac{1}{|C|}\sum_{i \in |C|} \delta_i$$

3.3 Robustness

This metric assesses the stability of the generated answers by evaluating their consistency under controlled input variations (Figure 2). We examine whether paraphrasing the original question leads to a similar response, and whether intentional changes to key parameters, such as hazard type and/or location, generate appropriately different responses. We consider two types of input modifications:

[3]See Appendix C.2 for details. We also consider token-level predictive entropy.

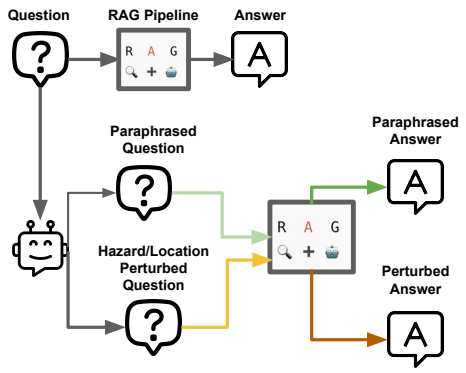


Figure 2: Robustness evaluation workflow. For each question–answer pair, the system generates paraphrased and hazard/location-perturbed variants of the question, runs them through the RAG pipeline, and compares the generated answers to assess semantic consistency and sensitivity to controlled perturbations.

Paraphrasing: We rephrase the original question while preserving its semantic meaning and pass the paraphrased question through the same RAG pipeline to generate a new answer. We then compute the cosine similarity between the original answer and the answer generated with the paraphrased question. A robust system should generate answers that are semantically consistent with the original response; therefore, higher cosine similarity indicates greater paraphrase robustness.

Perturbation: We modify the question by altering the hazard type and/or the geographical location mentioned. Unlike paraphrasing, these perturbations change the underlying query semantics. Therefore, we expect the system to retrieve a different set of relevant documents and produce correspondingly different answers. Similar to the paraphrasing setting, we pass the perturbed question through the same RAG pipeline and compute the cosine similarity between the original answer and the answer generated from the perturbed question. A lower cosine similarity indicates that the system is more sensitive to meaningful input perturbation.

3.4 Answer Relevance

Answer relevance evaluates whether a model's response is relevant to the original user query (Saad-Falcon et al., 2024; Es et al., 2024). We adapt the question-generation setup from the RAGAS framework (Es et al., 2024), where candidate questions are generated from the answer and compared with the original question. However, in our domain, high relevance scores can result from superficial lexical overlap, especially when domain-specific

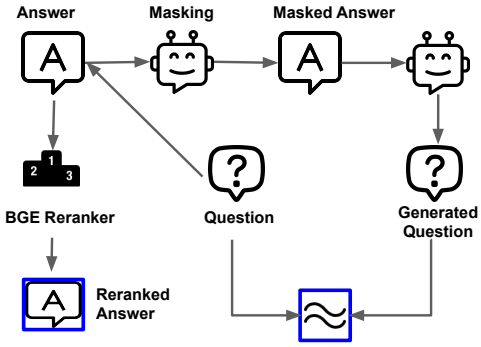


Figure 3: Answer relevance pipeline. For each answer, the system generates reranked answer and answer relevance score. Outputs are shown inside blue boxes.

keywords (e.g. specific hazard or infrastructure sector) appear in both the question and answer. To address this, we extend the RAGAS setup with a semantic masking step to capture conceptual alignment rather than surface-level overlap. Figure 3 shows the pipeline.

Answer Relevance with Masking. We apply a semantic masking step: an LLM first generates a masked answer $\tilde{a}$ by replacing sensitive tokens with a placeholder (e.g., replacing "electrical grid" with [INFRASTRUCTURE]), then generates $\tilde{q}$ from $\tilde{a}$ ensuring conceptual rather than lexical alignment. The masked relevance score is:

$$\tilde{\mathrm{R}}(q, a) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{sim}\,(q, \tilde{q}) \tag{3}$$

Relevance Scoring with BGE Reranker. To complement masked relevance, we apply a post-hoc reranking strategy to assess and enhance the informativeness of multi-point answers. We use a leave-one-out relevance attribution method, where each segment of the answer (e.g., bullet point or paragraph) is evaluated for its marginal contribution to the overall answer quality. We utilize the `BAAI/bge-reranker-base` model (Xiao et al., 2023), which provides a dense relevance score $R(q, a)$ between a question q and an answer a. A structured answer is usually composed of an optional introduction and m content segments $a := \text{Intro} \cup \{a_1, a_2, \ldots, a_m\}$. To estimate the marginal relevance of each individual segment a_i, we remove it from the answer to get A_{-i} and compute the difference:

$$g_i = R(q, a) - R(q, A_{-i}) \tag{4}$$

A higher g_i indicates a more significant contribution of that segment to the overall answer quality. If $g_i < 0$, it implies that the removed segment was detrimental to the overall relevance.

Finally, we rerank the answer segments $\{a_i\}$ in descending order of g_i, prepending the introduction (if present). This ensures that the most informative and relevant points appear earlier in the final answer presentation. Figure 4 shows an example.

4 Experimental Setup

We use LLMs as both answer generators within a RAG pipeline, and evaluators for the proposed metrics. We employ GPT-4o and Gemini 2.5 as the primary answer generators. Top five semantically relevant abstracts from our domain-specific knowledge base serve as grounding context per query. For evaluation, we use both proprietary and open-source models to show the framework's flexibility across resource tiers. Specifically, we employ GPT-4o as a commercial evaluator and Qwen3-8B as a representative open-source evaluator. We assess answer relevance by incorporating the BAAI/bge-reranker-base model to compute relevance scores for multi-point answers. Proprietary models like GPT-4o and Gemini do not expose token-level logits, needed to estimate CU score. We decouple answer generation and confidence estimation using teacher-forcing on lightweight models. Specifically, we use LLaMA3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Qwen3-8B as proxies and also experiment with Gemma2-9B-it and Ministral-8B-Instruct. See Appendix G for details.

4.1 Human Evaluation Setup

We recruit 10 annotators and assign each annotator 10 question–answer pairs through our dedicated annotation platform. To ensure reliability and measure inter-annotator agreement, every question is independently evaluated by two annotators. In total, the study covers 50 unique question–answer pairs. The annotation guidelines, platform details, and interface screenshots are provided in Appendix F.

5 Results

We present results based on our proposed SCORE framework using ClimateRiskQA, and show generalizability results for a clinically relevant data in Section 5.3.

Question Type	Generator	Evaluator	Specificity ↑	Robustness		AR ↑	CU ↑
				Para. ↑	Pert. ↓		
Fact-based	GPT-4o	GPT-4o/LLaMA (CU)	0.65 (0.19)	0.90 (0.05)	0.58 (0.09)	0.70 (0.12)	1.48 (0.84)
		Qwen3	0.62 (0.21)	0.89 (0.03)	0.65 (0.10)	0.74 (0.19)	1.60 (0.91)
	Gemini 2.5	GPT-4o/LLaMA (CU)	0.60 (0.21)	0.89 (0.05)	0.89 (0.06)	0.69 (0.11)	1.39 (0.92)
		Qwen3	0.46 (0.22)	0.87 (0.05)	0.65 (0.10)	0.83 (0.12)	1.46 (0.92)
Recommendation	GPT-4o	GPT-4o/LLaMA (CU)	0.56 (0.23)	0.89 (0.04)	0.60 (0.08)	0.69 (0.10)	1.82 (0.83)
		Qwen3	0.53 (0.23)	0.89 (0.03)	0.68 (0.09)	0.73 (0.15)	1.92 (0.92)
	Gemini 2.5	GPT-4o/LLaMA (CU)	0.52 (0.20)	0.86 (0.08)	0.86 (0.08)	0.68 (0.10)	1.38 (0.78)
		Qwen3	0.45 (0.34)	0.87 (0.04)	0.68 (0.09)	0.83 (0.11)	1.44 (0.84)
Hybrid	GPT-4o	GPT-4o/LLaMA (CU)	0.64 (0.19)	0.89 (0.07)	0.61 (0.09)	0.68 (0.12)	1.53 (0.94)
		Qwen3	0.61 (0.19)	0.89 (0.07)	0.64 (0.10)	0.74 (0.17)	1.67 (0.98)
	Gemini 2.5	GPT-4o/LLaMA (CU)	0.50 (0.20)	0.87 (0.09)	0.70 (0.04)	0.64 (0.08)	1.10 (0.65)
		Qwen3	0.42 (0.22)	0.86 (0.07)	0.63 (0.10)	0.81 (0.12)	1.10 (0.70)
Overall	GPT-4o	GPT-4o/LLaMA (CU)	**0.61 (0.21)**	**0.89 (0.05)**	**0.60 (0.09)**	0.69 (0.11)	1.63 (0.87)
		Qwen3	0.58 (0.22)	**0.89 (0.04)**	0.66 (0.10)	0.73 (0.18)	**1.74 (0.94)**
	Gemini 2.5	GPT-4o/LLaMA (CU)	0.55 (0.24)	0.87 (0.07)	0.87 (0.07)	0.68 (0.10)	1.34 (0.83)
		Qwen3	0.45 (0.27)	0.86 (0.05)	0.66 (0.10)	**0.83 (0.11)**	1.39 (0.87)

Table 1: Evaluation results for two different generators (GPT-4o and Gemini2.5) with mean and standard deviation of all measures for each generator-evaluator pair. Higher values indicate better performance for all metrics except perturbation robustness, where lower values show higher sensitivity to changes. Overall, GPT-4o produces more detailed and specific responses than Gemini 2.5. The evaluation results indicate that GPT-4o acts as a more lenient judge than Qwen3-8B.

5.1 Evaluation with SCORE Metrics

5.1.1 Specificity

Table 1 shows the specificity scores across generator-evaluator pairs. Across all settings, GPT-4o answers achieve higher specificity scores than Gemini, regardless of the evaluator. We also observe that specificity exhibits higher variance than other evaluation metrics. The higher variance in specificity scores suggests that model behavior is not uniform across questions. While GPT-4o often produces highly detailed answers, it occasionally generates less precise responses depending on the query. This variability reinforces the motivation for fine-grained specificity evaluation, as such detailed information is inherently difficult for models to provide accurately across diverse queries.

5.1.2 Context-Utilization

As shown in Table 1, GPT-4o exhibits consistently higher context utilization than Gemini 2.5 across all question types and evaluators, indicating that GPT-4o relies more heavily on the retrieved evidence when generating its answers. To test whether CU scores are sensitive to the choice of proxy model or the confidence measures, we report a sensitivity analysis in Table 2. LLaMA and Qwen variants exhibit consistent answer confidence when conditioned on retrieved documents D vs. claims C derived from it. This reflects consistency in the estimates as the claims C are simply a summary of D. In contrast, Gemma and Ministral show less reliable correlations. An extended analysis is reported in Appendix C.2 where LLaMA and Qwen variants also show near-perfect agreement in their CU scores and low agreement with Gemma and Ministral. Additionally, length-normalized geometric mean of token probabilities provides consistent confidence estimates than predictive entropy across models and generators (Appendix Table 9). Therefore, in Table 1 we report results with LLaMA and Qwen as evaluators using CU_{rel} and geometric-mean-based interpretation of f_θ. Mean CU_{rel} val-

Evaluator	$\rho(f_\theta(a\|q \cup C), f_\theta(a\|q \cup D)$	
	Generator=GPT-4o	Generator=Gemini2.5
LLaMA3.1-8B-Instruct	0.83 (0.0)	0.84 (0.0)
Gemma2-9B-it	0.17(4.4e-11)	0.12 (1.2e-05)
Ministral-8B-Instruct	0.14 (1.7e-07)	0.13 (6.0e-07)
Qwen2.5-7B-Instruct	0.95 (0.0)	0.96 (0.0)
Qwen3-8B	0.94 (0.0)	0.93 (0.0)

Table 2: Correlation between answer confidence conditioned on query q and all claims C vs. q and all documents D across evaluator models (p-values are reported in parentheses). Confidence f_θ is measured in terms of geometric mean of token probabilities. Correlations for Gemma and Ministral are insignificant, indicating their unsuitability for CU evaluation. In contrast, LLaMA and Qwen have significantly high positive correlation.

ues indicate that on average, removing a single claim reduces answer confidence by 1–2%. Detailed scores statistics are provided in Appendix Table 11. A small subset of claims causing substantially larger drops (9-10% CU_{rel}), indicating concentrated evidence dependence amid otherwise distributed evidence usage. Large negative values (e.g. CU_{rel} = -8.56%) highlight claims that harm generation. This demonstrates that the measure is effective at surfacing both highly supportive and detrimental evidence.

5.1.3 Robustness

We evaluate robustness under paraphrasing invariance and semantic perturbations. As shown in Table 1, both GPT-4o and Gemini exhibit high robustness to paraphrasing across all evaluator settings, which indicates that superficial linguistic changes do not significantly affect model outputs.

In contrast, semantic perturbations that alter key attributes such as hazard type or location lead to noticeable consistency drops. This behavior is desirable, as it indicates that models appropriately adjust their responses when the underlying facts change. Gemini shows higher perturbation robustness scores, which suggests that its responses change less under semantic modifications. This likely reflects a tendency toward more generic answers, which remain broadly applicable even when critical details are altered, whereas GPT-4o produces more detail-sensitive responses.

5.1.4 Answer Relevance

To evaluate how well responses address user queries, we use a dual strategy combining masked answer relevance and BGE reranking. As shown in Table 1, GPT-4o scored between 0.69 – 0.73, while Gemini 2.5 scores between 0.68 – 0.83. Additionally, using BGE reranker, we measure the value of each individual point in an answer by seeing how much the total score drops when that point is removed. Table 3 shows that Gemini achieves higher relevance under this metric. We also observe frequent reordering of answer points after reranking. It indicates that the relevance model meaningfully reweights the importance of individual answer points.

Higher relevance scores do not necessarily imply better answer quality: relevance captures semantic alignment with the query, whereas specificity reflects the presence of specific details. This motivates evaluating both metrics jointly.

Metric	GPT-4o	Gemini
Average BGE score	6.69	7.08
Has answer changed (yes)	1393	1397

Table 3: Comparison of BGE reranking for answers generated by GPT-4o and Gemini.

Gen.	Eval.	Spec.		AR		CU	
		Baseline	Ours	Baseline	Ours	Baseline	Ours
GPT-4o	GPT-4o/LLaMA3.1	0.83 (0.24)	0.65 (0.21)	0.85 (0.12)	0.64 (0.15)	0.89 (0.10)	1.75 (0.97)
	Qwen3	0.878 (0.225)	0.60 (0.23)	0.98 (0.091)	0.70 (0.17)	0.97 (0.12)	1.82 (1.07)
Gemini 2.5	GPT-4o/LLaMA3.1	0.76 (0.20)	0.55 (0.24)	0.93 (0.18)	0.70 (0.17)	0.87 (0.21)	1.30 (0.79)
	Qwen3	0.875 (0.189)	0.50 (0.25)	0.90 (0.19)	0.81 (0.14)	0.80 (0.34)	1.43 (0.83)

Table 4: Baseline evaluation results for specificity, answer relevance (AR), and context utilization (CU) across different generator–evaluator pairs. Values are reported as mean with standard deviation in parentheses.

5.2 LLM-as-a-Judge Baseline

We also test a direct LLM-as-a-judge baseline where an LLM rates each answer for specificity, relevance, and context utilization on a 0-1 scale. As shown in Table 4, the baseline assigns consistently high scores: 0.76–0.878 for specificity, 0.85–0.98 for answer relevance, and 0.80–0.97 for context utilization. In contrast, SCORE produces more varied metric-specific values, suggesting that the structured metrics capture finer differences. Qualitative case studies are provided in Appendix D.

5.3 Generalization to Clinical Domain

To test whether SCORE generalizes to other decision-critical domains beyond hazard response, we evaluate the framework on a clinically relevant subset of PubMedQA (Jin et al., 2019). Data processing details are provided in the Appendix E. As shown in Table 5, specificity is nearly perfect because PubMedQA answers are short and the provided contexts generally contain the relevant clinical details. Answer relevance is moderate, indicating that the answers address the biomedical questions but may omit some contextual details. The relative cu-score with LLaMA3.1 is reported in the table, indicating a low context utilization on average. We get similar results using Qwen3 as evaluator, with the mean relative CU score as -0.001, standard deviation 0.08, and observe small but significant correlation between Qwen3 and LLaMA3.1 scores ($\rho = 0.12, p = .041$). There's also perfect correlation in claim vs. context level confidence scores for both LLaMA and Qwen3. We analyzed several samples and find that generally the generated answers for PubMedQA are short and to the point, while context provides many extraneous details. Secondly, many additional numerical specifics (e.g.

Metric	Score
Specificity ↑	0.99 (0.06)
Answer Relevance ↑	0.63 (0.14)
Context Utilization ↑	-0.003 (0.05)

Table 5: Evaluation results for PubMedQA dataset. Scores are computed on GPT-4o-generated answers using GPT-4o/LLaMA (for CU) as the evaluator. Standard deviation are reported in parentheses.

Metric	Annotators	Anno-GPT	Anno-Qwen
Hazard	92	88	88
Location	86	77	73
Timeline	52	45	40
Intensity	78	75	59

Table 6: Agreement percentage between annotators and automated metrics for specificity.

exact percentages, patient counts) in the context were not utilized in the answer. Third, we also find some questions are ill-formed, leading to ambiguity in the answer. These issues are illustrated in the case studies presented in Appendix Figures 8 and 9. We do not report robustness for PubMedQA because robustness evaluation requires rerunning the full RAG pipeline with paraphrased and perturbed queries, whereas PubMedQA provides fixed question-context-answer triples.

5.4 Human Evaluation

We conduct human evaluation to measure inter-annotator agreement and alignment between human labels and automated evaluators. Because the task is open-ended and domain-specific, some disagreement is expected, especially for attributes requiring interpretation.

Inter-Annotator Agreement. We compute inter-annotator agreement across the 50 double-annotated questions. Table 6 shows agreement percentages (see Table 13 for details). Agreement is higher for concrete attributes such as hazard and location, and lower for subjective dimensions such as timeline, which often requires interpretation of vague temporal statements. For continuous metrics, we evaluate agreement using Spearman correlation. We obtain a Spearman correlation of 0.22 for answer relevance and 0.35 for context utilization, highlighting that both tasks are inherently subjective and cognitively demanding, with context utilization being slightly more stable than relevance.

Human vs. Automated Agreement. We compare human annotations with two automatic evaluators using row-level exact match on the same label ranges. Table 6 shows that alignment is generally higher for hazard and location, but drops for timeline and intensity (see Table 14 for details). We also notice that many human–automated mismatches occur on the same questions where annotators themselves disagree, which suggests that a large fraction of the errors come from genuinely ambiguous or subjective cases. For answer relevance, the Spearman correlations are 0.33 (human vs. GPT) and 0.32 (human vs. Qwen); for context utilization, they are 0.43 (for both human vs. Qwen and human vs. LLaMA). Matching or exceeding human–human agreement, our automated metrics appear as reasonable human-aligned proxies.

Error Analysis. Qualitative analysis shows that disagreements often arise when answers mention multiple hazards or use vague temporal expressions. In such cases, annotators and automated evaluators may differ on whether a detail is sufficiently specific or directly supported by the retrieved evidence. We provide detailed examples in Appendix F.

6 Conclusion

We introduce SCORE, a multi-dimensional evaluation framework for assessing LLM-generated answers in domain-specific, high-stakes settings, together with a new dataset designed to support evaluation where gold-standard references are scarce. By jointly analyzing specificity, context utilization, robustness, and answer relevance, we show that models exhibit distinct trade-offs between producing detailed, precise information and generating broadly relevant but more generic responses. Compared with a direct LLM-as-a-judge baseline, SCORE provides a more fine-grained and critical assessment by separately evaluating whether answers contain verifiable details, remain relevant to the query, and actually depend on the provided evidence. Our PubMedQA experiment further shows that the framework can generalize beyond climate-risk questions to another decision-critical domain by replacing only the domain-specific specificity schema. Our human evaluation highlights the challenges of aligning automated metrics with human judgment in open-ended, expert-level tasks, particularly for attributes requiring significant interpretation. These results emphasize the importance of evaluating LLM outputs using complementary metrics rather than a single score for a more reliable assessment of LLMs deployed in real-world decision support.

Ethical Considerations

This work focuses on the evaluation of large language model (LLM) outputs rather than the deployment of automated decision-making systems. Although our framework targets high-stakes domains such as natural hazard response and infrastructure planning, we emphasize that such systems should complement rather than replace human expert judgment, particularly in decisions affecting public safety and critical infrastructure. Our dataset is synthetically generated and contains no personal or personally identifiable information; however, while it is designed to reflect realistic hazard–location mappings, it may not capture the full complexity and diversity of real-world hazard scenarios, potentially disadvantaging regions or communities underrepresented in the climate and infrastructure literature. We also acknowledge that evaluation metrics themselves can introduce ethical risks if misinterpreted, as high scores on relevance, robustness, or specificity may create false confidence if treated as definitive indicators of system safety or deployment readiness rather than complementary assessment tools. Finally, all human annotators involved in this study were recruited and treated in accordance with institutional review board (IRB) guidelines, provided with clear information about the research purpose, participated voluntarily, and were compensated fairly for their time.

Limitations

The current framework relies on a specific knowledge base of approximately 600,000 climate-related records. A primary limitation is that if this knowledge base is insufficient or lacks coverage for a specific niche, the system may retrieve irrelevant documents. This can lead to error propagation, where the model generates an answer based on poor grounding. In addition, human evaluation of open-ended, domain-specific questions exhibits inherent subjectivity, particularly for attributes that require interpretation such as timeline and intensity.

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

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. Technical report.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. 2024a. Legalagentbench: Evaluating llm agents in legal domain. *arXiv preprint arXiv:2412.17259*.

Haobo Li, Zhaowei Wang, Jiachen Wang, Alexis Kai Hon Lau, and Huamin Qu. 2024b. Cllmate: A multimodal llm for weather and climate events forecasting. *arXiv preprint arXiv:2409.19058*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Kangcheng Luo, Quzhe Huang, Cong Jiang, and Yansong Feng. 2025. Automating legal interpretation with llms: Retrieval, generation, and evaluation. *arXiv preprint arXiv:2501.01743*.

Tanwi Mallick, Joshua David Bergerson, Duane R Verner, John K Hutchison, Leslie-Anne Levy, and Prasanna Balaprakash. 2023. Analyzing the impact of climate change on critical infrastructure from the scientific literature: A weakly supervised nlp approach. *arXiv preprint arXiv:2302.01887*.

Tanwi Mallick, Joshua David Bergerson, Duane R Verner, John K Hutchison, Leslie-Anne Levy, and Prasanna Balaprakash. 2025. Understanding the impact of climate change on critical infrastructure through nlp analysis of scientific literature. *Sustainable and Resilient Infrastructure*, 10(1):22–39.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223.

Aishik Nagar, Yutong Liu, Andy T Liu, Viktor Schlegel, Vijay Prakash Dwivedi, Arun-Kumar Kaliya-Perumal, Guna Pratheep Kalanchiam, Yili Tang, and Robby T Tan. 2024. umedsum: A unified framework for advancing medical abstractive summarization. *arXiv preprint arXiv:2408.12095*.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1890–1912.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467.

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

Yangxinyu Xie, Bowen Jiang, Tanwi Mallick, Joshua Bergerson, John K Hutchison, Duane R Verner, Jordan Branham, M Ross Alexander, Robert B Ross, Yan Feng, et al. 2025. Marsha: multi-agent rag system for hazard adaptation. *npj Climate Action*, 4(1):70.

A Additional Related Work

Prior work on reference-free LLM evaluation includes general-purpose LLM-as-a-judge benchmarks and RAG-specific evaluation frameworks. General-purpose benchmarks such as MT-Bench, AlpacaEval, and Chatbot Arena use high-capacity models or large-scale human preferences to evaluate conversational quality, instruction following, and pairwise model preference (Bai et al., 2024; Li et al., 2023; Chiang et al., 2024). These frameworks have been influential for comparing open-ended model outputs, but they are primarily designed for general conversational ability rather than domain-specific decision support.

RAG evaluation frameworks have introduced more fine-grained metrics for assessing generated answers conditioned on retrieved evidence. Methods such as RAGAS, ARES, RagChecker, and related claim-level verification approaches evaluate dimensions such as faithfulness, factuality, answer relevance, context relevance, and context utilization (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024; Min et al., 2023; Ni et al., 2024). These methods are useful for diagnosing whether generated answers are supported by retrieved documents and whether retrieved context is relevant to the query.

However, these frameworks do not fully address the needs of high-stakes natural hazard response. In this setting, an answer may be factually supported but still insufficient if it omits decision-critical details such as hazard type, location, timeline, intensity, infrastructure type, or user concern. Similarly, semantic answer relevance can be inflated by repeated domain keywords, even when the response remains generic. Context-overlap-based utilization can also overestimate evidence use because a model may produce claims that are consistent with retrieved context without actually relying on that context during generation. SCORE is designed to address these gaps by jointly evaluating specificity, robustness, answer relevance, and context utilization in a reference-free manner.

B Dataset Details

Here we provide details of the dataset generation.

B.1 Profession

Table 7 provides a summary 40 different professions across five sectors.

B.2 Hazard and Location

We select the location and hazard types based on realistic hazard-region mappings that frequently occur within the U.S. context. For instance, hurricanes are commonly associated with Miami or New Orleans, while wildfire scenarios with San Diego or Elko, Nevada. Table 8 summarizes the seven hazard types and their corresponding locations (counties and states) used in the question generation template.

B.3 Question Generation

Each question template contains placeholders for infrastructure, hazard, location, concern, and timeline, which are populated using the corresponding synthetic user profile attributes. We use three concern types: fact-based, recommendation-seeking, and hybrid. Fact-based templates focus on objective vulnerabilities, recommendation-seeking templates elicit mitigation or planning strategies, and hybrid templates capture infrastructure interdependencies and cascading impacts.

Examples of templates include:

- *What are the critical vulnerabilities of [INFRASTRUCTURE] to [HAZARD] in [LOCATION]?*
- *What strategies should [PROFESSION] consider to enhance [INFRASTRUCTURE] resilience in the face of [HAZARD] over the next [TIMELINE] years?*
- *What are the cascading impacts of [HAZARD] between [INFRASTRUCTURE] and building systems in [LOCATION]?*

During generation, the concern type in the user profile determines the template pool, from which one template is randomly selected.

C Evaluation Framework

In this section, we provide additional details on our evaluation framework.

C.1 Specificity

Atomic Claim Decomposition. We begin by decomposing each generated answer into a set of atomic claims, where each claim captures a single, self-contained factual statement. This decomposition enables us to evaluate the output at a more granular level, rather than treating the entire answer as one unit. By focusing on individual claims, we

Sector	Profession
Transportation	Highway Engineer, Bridge Inspector, Railway Systems Engineer, Transit Operations Manager, Airport Infrastructure Manager, Port Facility Manager, Transportation Safety Inspector, Traffic Systems Engineer, Pavement Engineer, Transportation Planner
Water	Water Systems Engineer, Hydraulic Engineer, Dam Safety Inspector, Wastewater Treatment Specialist, Maritime Infrastructure Manager, Stormwater Engineer, Water Quality Specialist, Coastal Infrastructure Engineer
Energy	Power Systems Engineer, Electrical Grid Manager; Energy Distribution Specialist, EV Infrastructure Planner, Renewable Energy Systems Manager, Transmission Line Engineer, Substation Engineer, Energy Storage Specialist
Buildings	Structural Engineer, Building Systems Manager, Facilities Manager, Real Estate Asset Manager, Building Automation Specialist, Construction Manager, Building Code Inspector, MEP Systems Engineer
Communications	Telecommunications Engineer, Broadband Infrastructure Specialist, Network Resilience Manager, Data Center Infrastructure Engineer, Fiber Optics Specialist, Communications Systems Planner, Network Security Engineer

Table 7: List of professions across five critical infrastructure sectors (Transportation, Water, Energy, Buildings, and Communications) used for generating synthetic user profiles in our dataset.

can perform more precise and targeted assessments of specificity. We can also identify whether each claim includes the fine-grained details that make it informative.

Specific Information Extraction. After decomposing each generated answer into claims, we extract the specific information that can be verified against the retrieved documents. In particular, we focus on four dimensions of specificity: *hazard type*, *location*, *timeline*, and *intensity*. To accomplish this, we prompt an LLM to identify and extract all explicit mentions of these dimensions from the claims. The extracted information is then passed to our specificity evaluation agents in the following step for verification and scoring.

Specificity Scoring. We evaluate each claim to determine whether it includes specific factual details that are supported by the knowledge base. The goal of this step is not to assess whether a claim is factually correct, but to determine whether fine grained information such as hazard type, location, timeline, or intensity is verified based on the knowledge base. We denote the input to the evaluation process as a triplet:(S, c_i, K), where S is specific information, c_i is the i-th atomic claim extracted from the generated answer, K is the corresponding knowledge base (e.g., abstracts from the literature). This triplet is provided to a language model evaluator agent, which is prompted to assess whether the claim c_i expresses concrete information that is *verifiable using the knowledge base* K, with respect to four pre-defined specificity dimensions:

$$\mathcal{D} = \{\text{hazard type, location, timeline, intensity}\}$$

Each agent produces a structured annotation for the claim across the four specificity dimensions from a fixed label set: $L = \{yes, no, n/a\}$. A label of "yes" indicates that the specific information is present in the claim and is also verifiable using the knowledge base. A label of "no" indicates that the information is not supported by the knowledge. Finally, "n/a" is used when a particular dimension does not apply to the claim's context (e.g., that specific information is not mentioned in the claim).

C.2 Context Utilization: Proxy-Based Scoring

We operate at the level of atomic claims rather than retrieved chunks. Claims are extracted from retrieved context using GPT-4o, prompted to identify unique, non-redundant factual propositions. This claim-level representation enables fine-grained attribution while avoiding redundancy within retrieved passages. To quantify each claim's contribution, we require an estimate of the model's confidence in generating a fixed answer given a particular set of claims. However, answers are often produced by closed or proprietary models (e.g., GPT or Gemini APIs) that do not expose token-level likelihoods. To enable reproducible and model-agnostic evaluation, we therefore decouple answer generation from confidence estimation. Concretely, for a fixed answer a, we compute confidence using teacher forcing on a lightweight open-source proxy language model. The proxy model is conditioned on the query q and context X, and token-level probabilities are estimated by forcing the model to follow the reference answer sequence.

Hazard Type	Counties and States
Coastal Flooding	Bergen, NJ; Atlantic, NJ; Ocean, NJ; Cape May, NJ; Hudson, NJ; Monmouth, NJ; Grays Harbor, WA; Middlesex, NJ; Kings, NY; Cumberland, NJ; Clatsop, OR; Cameron, TX; Philadelphia, PA; Queens, NY; Coos, OR; Bronx, NY; Sussex, DE; Westchester, NY; New York, NY; Jefferson, LA; Fairfield, CT; St. Charles, LA; Suffolk, NY; Aransas, TX; Union, NJ
Cold Wave	Cook, IL; Milwaukee, WI; Minnehaha, SD; Wayne, MI; Lake, IL; Nueces, TX; Lake, IN; Hennepin, MN; Williams, ND; Will, IL; Yakima, WA; Anoka, MN; Flathead, MT; Winnebago, IL; Pennington, SD; Dane, WI; Ramsey, MN; Cass, ND; Marathon, WI; Sheboygan, WI; Blue Earth, MN; Brown, WI; Olmsted, MN; Outagamie, WI; St. Louis, MN
Drought	Santa Barbara, CA; Yolo, CA; Sutter, CA; Napa, CA; Colusa, CA; Glenn, CA; Butte, CA; Sonoma, CA; Sacramento, CA; Solano, CA; Pinal, AZ; Floyd, TX; Lubbock, TX; Humboldt, NV; Doña Ana, NM; Maricopa, AZ; Yuma, AZ; Kings, CA; Imperial, CA; Merced, CA; Madera, CA; Stanislaus, CA; Fresno, CA; Tulare, CA; Kern, CA
Heat Wave	Cook, IL; Clark, NV; St. Louis, MO; Philadelphia, PA; Dallas, TX; Tulsa, OK; Maricopa, AZ; Queens, NY; Tarrant, TX; Kings, NY; Oklahoma, OK; Tulare, CA; Jackson, MO; Shelby, TN; Baltimore, MD; Fulton, GA; Los Angeles, CA; Harris, TX; Bexar, TX; Fairfax, VA; Franklin, OH; DeKalb, GA; Prince George's, MD; Mecklenburg, NC; Wayne, MI
Hurricane	Harris, TX; Miami-Dade, FL; Broward, FL; Palm Beach, FL; Hillsborough, FL; Lee, FL; Brevard, FL; Pinellas, FL; Charleston, SC; Pasco, FL; Horry, SC; Collier, FL; Chatham, GA; Mobile, AL; New Hanover, NC; Galveston, TX; Orange, FL; Volusia, FL; Indian River, FL; St. Lucie, FL; St. Johns, FL; Manatee, FL; Clay, FL; Beaufort, SC; Escambia, FL
Ice Storm	Nassau, NY; Tulsa, OK; Greene, MO; Lancaster, NE; St. Louis, MO; Oakland, MI; Boone, MO; Richland, SC; Monmouth, NJ; Washington, AR; Macomb, MI; Johnson, KS; Morris, TX; Baxter, AR; Rogers, OK; Douglas, NE; Sedgwick, KS; Linn, IA; Dubuque, IA; Stark, OH; Polk, IA; Peoria, IL; Knox, TN; Lucas, OH; Hamilton, OH
Wildfire	San Diego, CA; Riverside, CA; San Bernardino, CA; Los Angeles, CA; Washington, UT; Elko, NV; Ventura, CA; Orange, CA; Pima, AZ; Maricopa, AZ; Ravalli, MT; Kern, CA; Yavapai, AZ; Utah, UT; Madera, CA; Nevada, CA; Placer, CA; Shasta, CA; Siskiyou, CA; Tehama, CA; Santa Cruz, CA; Alameda, CA; Tuolumne, CA

Table 8: Hazard types and the corresponding counties and states included in our dataset generation framework.

This yields a consistent likelihood estimate without requiring access to the original model's internal probabilities.

Sensitivity Analysis of Proxy Models. A natural concern is whether the CU scores are sensitive to the choice of proxy models or confidence measures. To address this, we conduct a sensitivity analysis and measure:

1. **Confidence Consistency** as the correlation $\rho(f_\theta(a|q \cup C), f_\theta(a|q \cup D))$ between answer confidence given query q and all claims C and q and the original retrieved documents D.

2. **CU Agreement**: correlation in CU scores across proxy models.

Table 9 reports Confidence Consistency—correlation between answer confidence when conditioned on query q and claims C ($f_\theta(a|q \cup C)$) extracted from retrieved information D compared to conditioning on query and D. These are computed based on both our confidence measures—geometric mean of token-level probabilities and predictive entropy. If a proxy model's token-level probabilities are well calibrated, the two would be highly correlated, since C summarizes D. LLaMA and Qwen variants exhibit high confidence consistency, while Gemma and Ministral show substantially lower correlations, suggesting they might be less suitable for proxy-based CU estimation. Across models, normalized log-likelihood yields more consistent confidence estimates than predictive entropy.

Further, Table 10 reports **CU Agreement** as the pairwise correlation in CU and CU_{rel} across proxy models. CU scores from LLaMA and Qwen variants are nearly perfectly correlated, whereas correlations involving Gemma and Ministral are lower, although they remain positive. These patterns are consistent for answers generated by both GPT-4o and Gemini-2.5.

11 provides detailed statistics of the CU-scores. The mean CU values are small and positive, indi-

cating that, on average, removing individual claims leads to a modest but consistent drop in answer confidence. Negative minimum values reflect cases where removing a claim increases confidence, indicating distracting or redundant context—a behavior expected in realistic retrieval settings. Importantly, these negatives are rare and do not affect the overall positive mean, implying that most retrieved claims contribute constructively. Relative CU values provide a more interpretable signal and is therefore our primary metric.

C.3 Answer Relevance

Figure 4 also shows the BGE reranking component used in our answer relevance evaluation. For each multi-point answer, we remove one segment at a time and measure the resulting change in the BGE relevance score. This leave-one-out attribution estimates each segment's contribution to overall answer relevance and allows us to rerank the answer so that the most informative points appear earlier.

C.4 Readability

The next metric we show is readability. Since our questions originate from domain-specific, graduate-level, or professional contexts, we expect the model-generated answers to match an appropriate level of linguistic complexity. To evaluate this, we employ two widely used readability measures. The first is the Flesch Reading Ease (FRE) score (Flesch, 1948), a numerical metric ranging from 0 to 100+, where higher values indicate easier text. The FRE score is defined as:

$$\text{FRE} = 206.835 - 1.015 \cdot \text{ASL} - 84.6 \cdot \text{ASW} \quad (5)$$

where ASL denotes the average sentence length (words per sentence), and ASW represents the average number of syllables per word. According to conventional interpretation, scores above 60 correspond to plain English that is understandable to the general public (approximately 8th–9th grade reading level), while scores below 30 indicate highly complex or academic text, typically suitable for college-level readers or above.

The second metric we consider is the Flesch–Kincaid Grade Level (FKGL) (Kincaid et al., 1975), which maps text complexity directly onto a U.S. school grade level. FKGL is defined as:

$$\text{FKGL} = 0.39 \cdot \text{ASL} + 11.8 \cdot \text{ASW} - 15.59 \quad (6)$$

A score of 8.0 indicates that the text is readable by an average eighth-grade student, whereas a score of 16.0 or higher suggests that graduate-level reading proficiency is required.

We apply these metrics to evaluate and compare the readability of responses generated by two prominent large language models: **GPT-4o** and **Gemini**. For each generated response, we compute the FRE and FKGL scores and categorize them into standard readability levels. These include very easy, plain English, college level, and graduate level. Similarly, FKGL scores are grouped into categories such as elementary, middle school, high school, undergraduate, and graduate reading levels.

A summary of this analysis is presented in Table 12. This includes average FRE and FKGL scores for each model, as well as the distribution of their outputs across interpretive readability levels. These results allow us to assess not only the average complexity of model outputs, but also how frequently they produce text that is potentially inaccessible to a general audience.

C.5 LLM-as-a-judge Baseline

Table 4 shows the baseline scores for specificity, answer relevance, and context utilization. The baseline tends to produce high scores (e.g., AR and CU often > 0.9 for both generators and evaluators), making it difficult to distinguish between answers. In contrast, SCORE yields a wider and more interpretable spread of scores across models and dimensions. These results indicate that while a simple LLM-as-a-judge baseline is feasible, it behaves generically by assigning uniformly high scores, whereas SCORE provides more discriminative, structured, and interpretable evaluations. We present a few case studies illustrating these in Figures 5-7.

D Case Studies

We provide qualitative case studies to illustrate how the proposed metrics capture complementary aspects of answer quality beyond numerical scores. Figure 5 presents a specificity case study where LLM-as-a-Judge baseline results in high score ignoring fine-grained details. This illustrates how SCORE distinguishes detailed, responses from generic answers that may appear relevant but lack key information. Figure 6 presents a high context-utilization case study and Figure 7 presents a low context-utilization case study.

Evaluator	Generator=GPT-4o		Generator=Gemini2.5	
	(f_θ=log-prob)	(f_θ=entropy)	(f_θ=log-prob)	(f_θ=entropy)
LlaMA3.1-8B-Instruct	0.83 (0.0)	0.40 (5.7e-55)	0.84 (0.0)	0.58 (2.1e-128)
Gemma2-9B-it	0.17(4.4e-11)	0.16(4.2e-09)	0.12 (1.2e-05)	0.19 (2.1e-13)
Ministral-8B-Instruct	0.14 (1.7e-07)	0.13 (2.0e-06)	0.13 (6.0e-07)	0.38 (3.2e-50)
Qwen2.5-7B-Instruct	0.95 (0.0)	0.46 (3.9e-76)	0.96 (0.0)	0.62 (5.8e-150)
Qwen3-8B	0.94 (0.0)	0.49 (1.7e-85)	0.93 (0.0)	0.60 (5.3e-137)

Table 9: Correlation (ρ) between $f_\theta(a|q \cup C)$ and $f_\theta(a|q \cup D)$. Where C denotes all extracted claims from retrieved content R. P-values are reported in parentheses. The correlations for Gemma and Ministral are positive but low, indicating the token probabilities might not be well-calibrated.

Evaluators	Generator=GPT-4o		Generator=Gemini2.5	
	Δ	δ	Δ	δ
LLaMA vs. Gemma	-0.00 (0.87)	-0.09 (0.0)	-0.10 (2.94e-04)	-0.09 (3.87e-04)
Ministral vs. Gemma	0.66 (1.2e-177)	0.01 (0.63)	0.45 (1.2e-69)	0.01 (0.68)
Ministral vs. LLaMA	0.03 (0.25)	0.06 (0.02)	-0.02 (0.48)	0.01 (0.69)
Ministral vs. Qwen2.5	0.00 (0.87)	0.06 (0.04)	-0.01 (0.7)	0.05 (0.06)
Qwen2.5 vs. LLaMA	0.87 (0.0)	0.86 (0.0)	0.84 (0.0)	0.80 (2.5e-321)
Qwen2.5 vs. Gemma	-0.02 (0.42)	0.01 (0.77)	-0.10 (1.7e-04)	-0.04 (.13)
Ministral vs. Qwen3	-0.03 (0.32)	0.05 (0.04)	-0.02 (0.41)	0.02 (0.43)
Qwen3 vs. LLaMA	0.81 (0.0)	0.83 (0.0)	0.85 (0.0)	0.79 (1.2e-295)
Qwen3 vs. Gemma	-0.03 (0.30)	0.00 (1.00)	-0.08 (.002)	-0.03 (0.2)
Qwen3 vs. Qwen2.5	0.82 (0.0)	0.83 (0.0)	0.83 (0.0)	0.82 (0.0)

Table 10: Pairwise correlation between context utilization scores from different proxy models using both absolute (Δ) and relative (δ) scores. P-values are reported in parentheses.

Generator	**Evaluator**	CU			$CU^{\%}_{rel}$		
		min	max	mean	min	max	mean
GPT-4o	LLaMA3.1-8B-Instruct	-0.001	0.024	0.004 (0.002)	-0.17	9.82	1.63 (0.87)
	Qwen3-8B	0.000	0.019	0.004 (0.002)	-0.51	9.63	1.74 (0.94)
Gemini2.5	LLaMA3.1-8B-Instruct	-0.001	0.014	0.003 (0.002)	-8.56	6.45	1.34 (0.83)
	Qwen3-8B	-0.001	0.015	0.003 (0.002)	-0.74	7.35	1.39 (0.87)

Table 11: CU scores across two generators and two evaluators (standard deviation is in parentheses), showing on average modest but positive contributions from individual claims to answer confidence, with rare distracting claims. Relatively larger max values indicate concentrated evidence dependence.

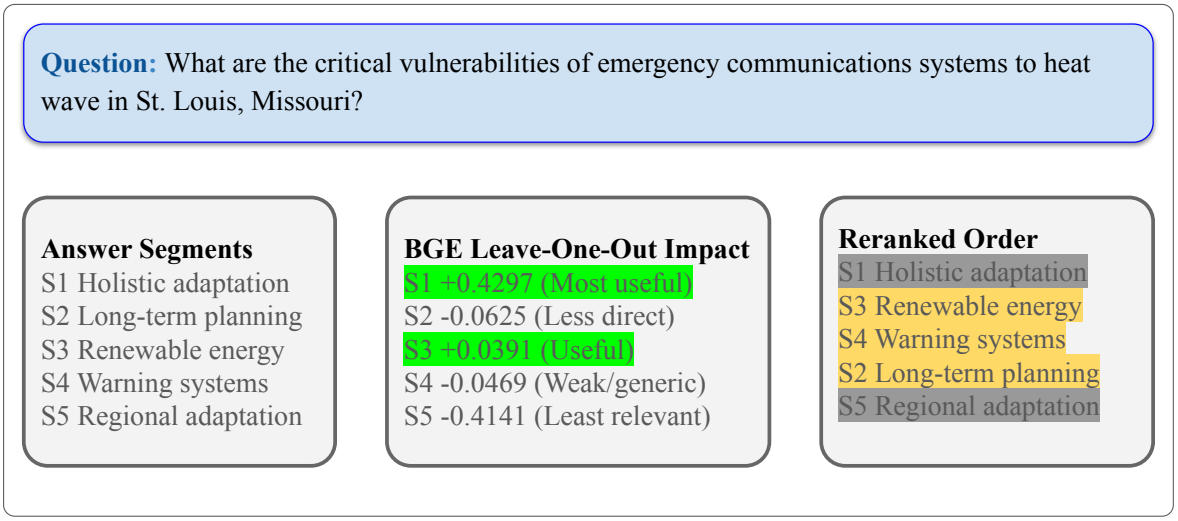


Figure 4: Example of BGE answer reranking. Each answer segment is removed and the change in the relevance is used to estimate the marginal contribution. Segments with positive impact improve relevance and are ranked earlier, while segments with negative impact are less direct, generic, or potentially distracting.

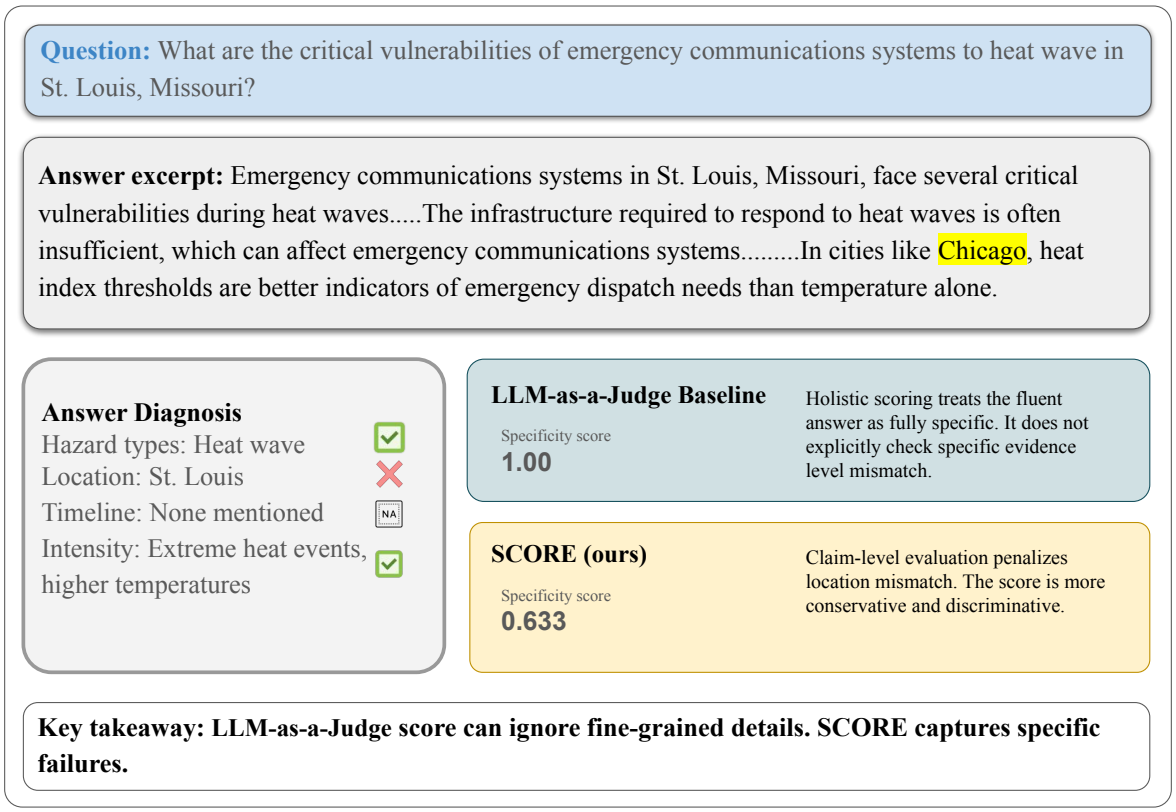


Figure 5: Comparison between a direct LLM-as-a-judge baseline and SCORE for specificity evaluation. The baseline LLM assigns a perfect score to a fluent answer, while SCORE performs claim-level and dimension-specific checking and produce a more conservative and discriminative specificity score.

Here is my profile:
Profession: Real Estate Asset Manager
Concern: Hybrid
Location: Washington, Utah
Timeline: 10 - 20 years
Scope: wastewater system wildfire

What are the cascading impacts of wildfire between wastewater system and building systems in Washington, Utah?

CU-SCORE

Llama: **6.39**
Qwen3: **8.41**

LLM-AS-JUDGE

GPT-4o: **0.7**

Matched in both | Context unused in answer | (1) Matching pair number

CONTEXT CLAIMS

The impacts of a September 2010 wildfire on hydrologic and biogeochemical processes are being evaluated in a Colorado Front Range stream (1)

A report focuses on options for protecting structures and for protecting wildlands and natural resources from wildfires. (2)

Fuel reduction treatments are often designed to achieve multiple resource management objectives in addition to reducing potential fire hazard. (3)

ANSWER

1. As a real estate asset manager in Washington, Utah, understanding the cascading impacts of wildfires on wastewater and building systems is crucial for long-term planning and asset protection. The research abstracts provide insights into how wildfires affect hydrologic processes, which can directly impact wastewater systems. The study on the Colorado Front Range stream highlights that wildfires can alter hydrologic and biogeochemical processes, potentially affecting water quality (1) and availability for wastewater systems (Evaluating the effects of wildfire on stream processes in a Colorado front range watershed, USA).

2. The conceptual model from the 1995 Arizona conference suggests that prescribed fires can be used to manage watershed resources, which may mitigate some negative impacts of wildfires on wastewater systems. This approach could be considered in your long-term strategy to protect both natural resources and infrastructure (Effects of Prescribed Fire on Watershed Resources: A Conceptual Model).

3. The report on wildfire damages emphasizes the importance of protecting structures and natural resources from wildfires. It discusses fuel treatment options, which can reduce fire hazards and potentially limit damage to building systems (3) and wastewater infrastructure. Implementing these treatments could be beneficial in your 10-20 year timeline to safeguard assets (Wildfire Damages to Homes and Resources: Understanding Causes and Reducing Losses).

4. The simulation study from the 1987 Arizona conference provides insights into the hydrologic impacts of fire, which are relevant for understanding how wildfires might affect wastewater systems. These impacts can include changes in water flow and quality, which are critical considerations for maintaining effective wastewater management (Simulating the Impacts of Fire: A Hydrologic Component).

5. Fuel treatments, as discussed in the study on forest structure and fire severity (3) in Arizona, can change the spatial patterns of fire severity. This can be strategically used to protect building systems and wastewater infrastructure by reducing the intensity and spread of wildfires in vulnerable areas (Fuel treatments change forest structure and spatial patterns of fire severity, Arizona, USA).

Confidence Score: 85% Explanation: The recommendation is based on relevant literature that addresses the user's concern about the impacts of wildfires on wastewater and building systems. While the abstracts provide useful insights, they are not specific to Washington, Utah, which slightly limits the applicability of the findings. However, the principles discussed are broadly applicable and can inform strategic planning in the user's context.

Figure 6: Comparison between a direct LLM-as-a-judge baseline and SCORE on an example with high context utilization. Despite strong use of retrieved context, the baseline assigns only above average score 0.7 (range 0-1) similar to lower-utilization case in figure 7 where it assigns 0.6, whereas SCORE (ours) better reflects the degree of contextual grounding making this among the highest scoring examples.

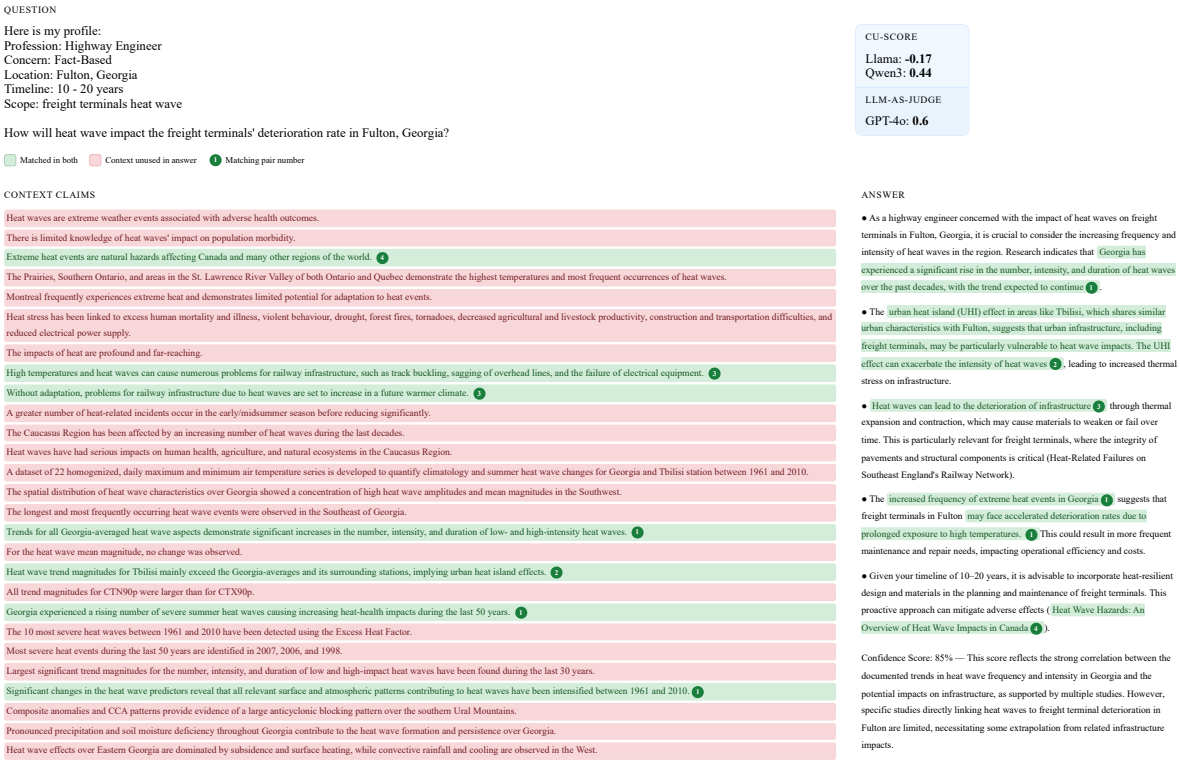


Figure 7: Comparison between a direct LLM-as-a-judge baseline and SCORE on examples with low context utilization. The baseline remains relatively insensitive to reduced context usage, while SCORE provides a more differentiated assessment based on claim level confidence check, making this among the lowest scoring examples.

Metric	GPT-4o	Gemini
Average Flesch Reading Ease	12.84	10.98
Average Flesch-Kincaid Grade Level	17.80	18.22
#Fairly difficult (10th–12th grade)	0	2
# Difficult to read (college level)	37	93
# Very difficult (college graduate or higher)	1375	1317
#Middle school level	1	2
#High school level	0	28
# College undergraduate level	167	295
# Graduate/professional level	1244	1087

Table 12: Comparison of automated readability scores for answers generated by GPT-4o and Gemini.

These examples show that direct baseline scoring is less discriminative, often assigning similarly high scores across responses, whereas SCORE captures example-specific differences.

E Generalizability Setup

To construct the PubMedQA subset, we use an LLM-based filtering to identify clinically relevant questions. A question is retained if it could plausibly be used by a physician or healthcare practitioner to consult biomedical literature for clinical understanding or decision-making. We include questions involving diseases or disorders, treatments or procedures, diagnostic tests, measurements or biomarkers, and patient populations and exclude questions that are focused on basic molecular biology, protein/gene interactions, taxonomy or genomics-only questions. The dataset contains short yes/no/maybe labels and author-written long answers as the ground truth, along with retrieved contexts as evidence. Since our goal is to evaluate LLM-generated long answers in a RAG-based setting, we use the provided context and question to generate an answer with an LLM (without using the ground-truth answers). For specificity eval-

QUESTION

Are fascial closure after open abdomen: initial indication and early revisions decisive factors — a retrospective cohort study?

Matched in both | Context unused in answer | (1) Matching pair number

CU-SCORE

Llama: 0.17
Qwen3: 0.02

CONTEXT CLAIMS

- The surgical treatment method in which the peritoneal cavity is opened anteriorly and deliberately left open is often called 'open abdomen'.
- Open abdomen has become the standard of care in damage-control procedures.
- Open abdomen is used in the management of intra-abdominal hypertension.
- Open abdomen is used in the management of severe intra-abdominal sepsis.
- Traditionally, open abdomen has been closed in two stages.
- A modern trend is to close the fascial layers within the initial hospitalization to avoid complications. (1)
- Complications of open abdomen include enterocutaneous fistula and hernia formation.
- The aim of this study was to determine crucial factors influencing the possibility of fascial closure after open abdomen. (2)
- Between 2003 and 2013, 355 adult patients were treated with open abdomen in the institution.
- Patients were divided into two groups depending on fascial closure or not.
- Fascial closure group had 137 patients (39%) and non-fascial closure group had 218 patients (61%).
- Patients who reached fascial closure had a significantly higher rate of initially performed open abdomen. (3)
- The rate of initially performed open abdomen in the fascial closure group was 71%. (2)
- The rate of initially performed open abdomen in the non-fascial closure group was 54%. (2)
- The periods of time until a second look operation were significantly shorter in the fascial closure group. (4)
- The periods of time until a third look operation were significantly shorter in the fascial closure group. (4)
- The presence of peritonitis was significantly higher in the fascial closure group. (5)
- The presence of peritonitis in the fascial closure group was 47%.
- The presence of peritonitis in the non-fascial closure group was 38%.
- Large bowel resection was significantly higher in the fascial closure group. (6)
- The presence of large bowel resection in the fascial closure group was 54%.
- The presence of large bowel resection in the non-fascial closure group was 41%.
- Rates of in-hospital mortality were significantly higher in the non-fascial closure group. (7)
- In-hospital mortality in the non-fascial closure group was 28%.
- In-hospital mortality in the fascial closure group was 44%.
- The presence of pancreatitis was significantly higher in the non-fascial closure group. (8)
- The presence of pancreatitis in the non-fascial closure group was 9%.
- The presence of pancreatitis in the fascial closure group was 2%.

ANSWER

Final decision: YES

The study indicates that initial indication and early revisions are decisive factors for fascial closure after open abdomen (1), confirming the study's aim to identify crucial influencing factors. (2)

Patients who achieved fascial closure had higher rates of initially performed open abdomen (71% vs. 54% in the non-closure group) (3), suggesting that the original surgical indication plays a significant role in whether fascial closure can ultimately be achieved.

Shorter times to subsequent second and third look operations were observed in the fascial closure group (4), indicating that early surgical revisions are associated with successful closure outcomes.

Additionally, certain clinical factors were significantly more prevalent in the fascial closure group: the presence of peritonitis was higher in the fascial closure group (5) and large bowel resection was more common among patients who achieved closure (6).

Conversely, adverse outcomes were concentrated in the non-closure group: in-hospital mortality rates were significantly higher among patients who did not achieve fascial closure (7), and the presence of pancreatitis was significantly more frequent in the non-fascial closure group (8).

Figure 8: High CU-score example from PubmedQA.

QUESTION

Is breathlessness at rest the dominant presentation of patients admitted with heart failure?

Matched in both | Context unused in answer | (1) Matching pair number

CU-SCORE

Llama: -0.20
Qwen3: -0.02

CONTEXT CLAIMS

- Many assume that most patients hospitalized with HF are short of breath at rest (SOBAR). (1)
- The National HF Audit suggests that this assumption is false. (1)
- A retrospective case-note review was carried out of patients hospitalized with HF.
- The review determined how many present with shortness of breath at rest or are comfortable at rest but breathless on slight exertion (CARBOSE). (3)
- Vital signs were tracked for 24 h and mortality for 180 days.
- Of 311 patients, those who were SOBAR (42%) had higher median heart rate (100 vs. 85 b.p.m.; $P < 0.001$). (2)
- Those who were SOBAR had higher systolic blood pressure (SBP) (141 vs. 122 mmHg; $P < 0.001$).
- Those who were SOBAR had higher RR (24 vs. 18 breaths/min; $P < 0.001$) compared with those who were CARBOSE (56%). (3)
- Vital signs changed little in those who were CARBOSE over the first 4–6 h.
- SBP fell in those who were SOBAR from 141→128 mmHg ($P < 0.001$).
- HR fell in those who were SOBAR from 100→90 b.p.m. ($P = 0.002$).
- RR fell in those who were SOBAR from 24→20 breaths/min ($P < 0.001$).
- At presentation, SBP >125 mmHg in 73% of patients who were SOBAR.
- At presentation, SBP >125 mmHg in 46% of patients who were CARBOSE.
- SBP dropped to 52% in SOBAR and 37% in CARBOSE by 4–6 h.
- Mortality amongst those who were SOBAR was 19%.
- Mortality amongst those who were CARBOSE was 34%.
- Odds ratio for mortality between SOBAR and CARBOSE was 2.29 ($P = 0.005$, 95% CI 1.29–4.06).

ANSWER

Final decision: YES

Breathlessness at rest is not the dominant presentation of patients admitted with heart failure (1), as only 42% of patients were short of breath at rest (2), compared to 56% who were comfortable at rest but breathless on slight exertion (3)

Figure 9: Low CU-score example from PubmedQA.

uation, we focus on four key biomedical dimensions—disease or disorder, treatment or procedure, population, and measurement/test/biomarker. Unlike the natural hazard setup, where answers are decomposed into atomic claims, we skip the claim decomposition step for PubMedQA because the generated answers are typically short and often consist of a single concise response. We then apply the remaining specificity scoring procedure directly to the full answer: biomedical details are extracted, three evaluator agents assign yes/no/N/A labels for each dimension, and the final specificity score is computed by aggregating valid dimension-level judgments. We use same setup as before for the other metrics.

F Human Evaluation

We recruit 10 annotators to validate our evaluation scores. Each annotator first logged into the evaluation platform and was presented with a personalized dashboard listing their assigned questions (see Figure 10). The dashboard clearly enumerates each question (e.g., Q16, Q21, Q1, etc.) along with an "Annotate" button that opens the detailed evaluation page. Upon selecting a question, annotators were shown a structured evaluation interface (see Figure 11) that displayed the user profile used to generate the answer alongside the question itself, the model-generated answer (formatted in numbered points following our prompt design), and the five retrieved knowledge sources that informed the response. Annotators then evaluated each answer across four dimensions: Specificity, which included sub-criteria for hazard type match, location match, timeline match, and intensity match; Answer Relevance, rated on a 1–10 scale; Context Utilization, which assessed whether the retrieved documents were actually used in the generated answer; and Overall Confidence, where annotators reported how confident they were in their evaluation (1–10) and could optionally provide comments. Each metric included clear guidance indicating when an annotator should select Yes, No, or N/A, along with radio-button options to standardize responses.

Agreement. Table 13 shows agreement, disagreement and agreement percentage between human annotators over 50 double-annotated question–answer pairs. Table 14 shows the same metrics for agreement between human labels and automated evaluators. Note that each question is annotated by two annotators, yielding 100 question-answer pairs.

Metric	Agree	Disagree	Agree (%)	Fleiss' kappa
Hazard	46	4	92	0.6678
Location	43	7	86	0.6834
Timeline	26	24	52	0.2640
Intensity	39	11	78	0.4579

Table 13: Inter-annotator agreement across 50 double-annotated questions. Agreement is higher for more concrete attributes (e.g., hazard, location) and lower for more subjective ones (e.g., timeline).

Metric	Agree		Disagree		Agree (%)	
	GPT	Qwen	GPT	Qwen	GPT	Qwen
Hazard	88	88	12	88	88	88
Location	77	73	23	27	77	73
Timeline	45	40	55	60	45	40
Intensity	75	59	25	41	75	59

Table 14: Human vs. automated evaluator agreement over 100 human-annotated rows (50 questions × 2 annotators), computed as row-level exact match. Alignment is higher for hazard and location and lower for timeline, intensity, and answer relevance.

Error Analysis. We conduct a qualitative error analysis to understand common failure modes and sources of disagreement. For specificity-hazard, some answers discuss multiple hazards (e.g., ice storms, storm surge, sea level rise) in the same response. In these cases, some annotators score hazard specificity only with respect to the hazard explicitly asked in the question, while others treat any hazard details mentioned in the answer as sufficient. For specificity-timeline, we observe boundary cases where the answer mentions a timeline and the retrieved sources also include temporal information, but the timelines do not match. In one example, the answer claims impacts over the "next 30–50 years," but that exact range is not stated in the retrieved evidence. The human annotators marked *Yes* because a timeline was mentioned, GPT marked *N/A* due to vague temporal framing, and Qwen marked *No* because the "30–50 years" claim was not verifiable from the sources. Figure 12 shows one representative example, including the question, context, retrieved knowledge (partial), and the generated answer. In this case, the answer includes temporal information, but it is stated in a vague form compared to the evidence. The retrieved sources discuss time-related effects, but they do not support that exact range. This creates an ambiguity: one may mark the timeline as supported because a timeline is mentioned, while a stricter

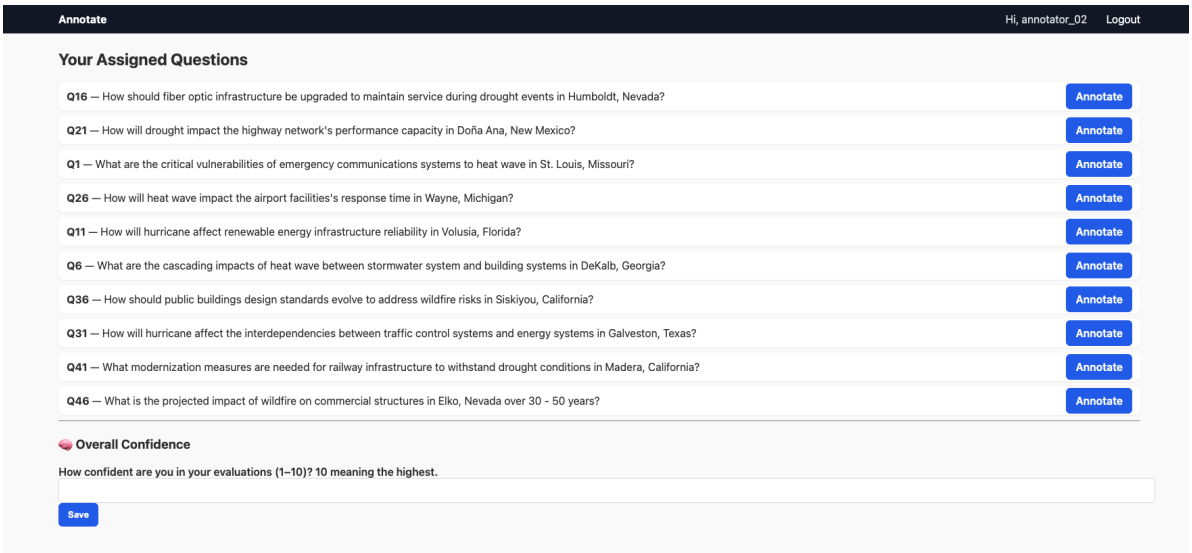
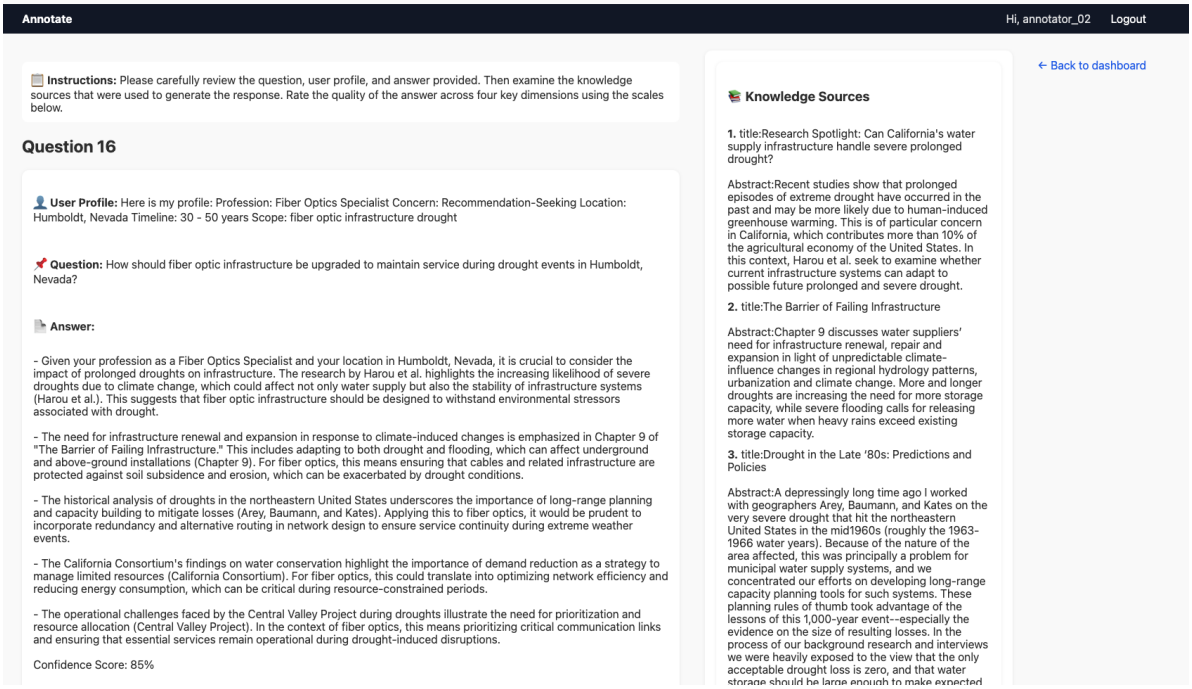


Figure 10: The annotation dashboard where each annotator receives ten questions and evaluates them by selecting the "Annotate" button for each item.

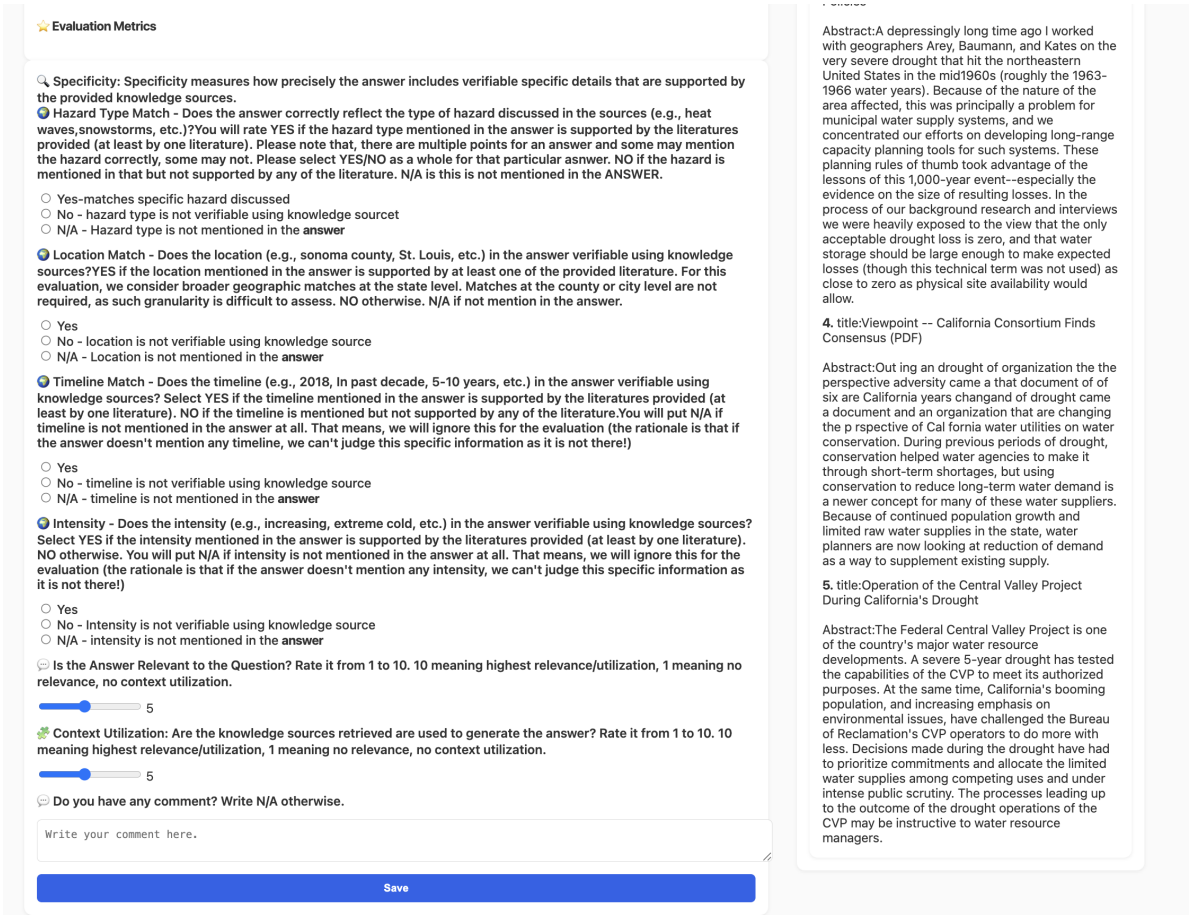
interpretation marks it as unsupported because the claimed timeline is not explicitly verifiable from the provided evidence.

G Implementation Details

All experiments are conducted using fixed decoding and evaluation configurations to ensure consistency across both proprietary and open-source models. For answer generation, we use a temperature of 0.1 and a top-p of 0.9 for both GPT-4o and Gemini-2.5-flash to prioritize factual precision and minimize stochastic variation, which is particularly important in high-stakes domains. For literature retrieval, we use a semantic search system built on a knowledge base containing approximately 600,000 climate-related records. Query embeddings are generated using the `all-MiniLM-L6-v2` SentenceTransformer model and indexed with FAISS (Facebook AI Similarity Search). For each query, we retrieve the top five relevant literature abstracts ($k = 5$), which are used as the grounding context for all generated responses. For robustness metric, we compute cosine similarity over sentence-level embeddings obtained from `all-MiniLM-L6-v2` SentenceTransformer model. All proprietary models are accessed via APIs. Local model inferences for LLaMA3.1-8B, Qwen3-8B etc., are performed using the vLLM framework on an NVIDIA A100 (80GB) GPU to ensure stable, high-throughput, and efficient inference.



(a) Annotation Interface – Part 1



(b) Annotation Interface – Part 2

Figure 11: Human Annotation Interface of our evaluation pipeline.

Question: What are the cascading impacts of ice storm between coastal protection and building systems in Richland, South Carolina?

Context: Profession: Wastewater Treatment Specialist, Concern: Hybrid, Location: Richland, South Carolina, Timeline: 30 - 50 years

Answer: As a Wastewater Treatment Specialist in Richland, South Carolina, your concern about the cascading impacts of ice storms on coastal protection and building systems is well-founded, especially given the potential for climate change to exacerbate these issues over the next 30 to 50 years. The research indicates that cold regions, which cover a significant portion of the United States, experience engineering impacts from ice-related phenomena such as freeze-thaw cycles and ice jams. These can lead to structural damage, increased flood fighting costs, and environmental damage (U.S. Army Corps of Engineers).
- Although Richland, South Carolina, is not traditionally considered a cold region, the increasing unpredictability of weather patterns due to climate change could lead to more frequent and severe ice storms. This is supported by findings that climate change may exacerbate problems in temperate regions, leading to increased risk and uncertainty from ice jams and related flooding (Investigating the effects of long-term hydro-climatic trends on Midwest ice jam events).

Retrieved-Knowledge: Cold regions cover more than half the United States. Engineering impacts of cold regions include those caused by temperature (e.g., freeze–thaw cycling, atmospheric icing), river and lake ice (e.g., ice jams, operational impacts of ice), and snow. Damages include the potential The City of Iqaluit, Nunavut, is an expanding urban centre with important infrastructure located in the coastal zone. This study investigates the exposure of this infrastructure to coastal hazards (rising mean sea level, extreme water levels, wave run-up, and sea ice).....
Using a coastal digital elevation model, we evaluate the inundation and flooding that may result from projected sea level rise. Some public and private infrastructure is already subject to flooding during extreme high water events. Using a near upper-limit scenario of 0.7 m for relative sea level rise from 2010 to 2100, we estimate that critical infrastructure will have a remaining freeboard of 0.3–0.8 m above high spring tide, and some subsistence infrastructure will be inundated. The large tidal range, limited over-water fetch, and wide intertidal flats reduce the risk of wave impacts. When present, the shorefast ice foot provides protection for coastal infrastructure. The ice-free season has expanded by 1.0–1.5 days per year since 1979, increasing the opportunity for storm-wave generation and thus exposure to wave run-up. Overtopping of critical infrastructure and displacement by flooding of subsistence infrastructure are potential issues requiring better projections of relative sea level change and extreme high water levels.....
Hurricane Irene made landfall as a sizeable Category I storm at the outer banks of North Carolina on August 27, 2011. Overland wind gusts associated with the storm measured nearly 115 MPH, and tide surge reached nearly 10 feet. The storm impacted numerous communities along the coast line and sounds, and it also damaged many communities and structures miles upstream in several larger river basins. This paper presents a summary of damages attributed to wind and storm surge for approximately 30 dwellings dispersed across eastern North Carolina.....

Figure 12: Example from the error analysis showing the question, context, retrieved evidence, and generated answer. The highlighted text shows temporal information.

H Prompts

H.1 Answer Generation

For generating answers, we provide five retrieved documents, user profile and question in the prompt.

Prompt for Answer Generation

You are tasked with writing a recommendation/fact-based answer that answers the user's question based on a provided list of research abstracts and contextual information. Your response must:

1. Directly address the user's concern, ensuring the answer is supported by the provided literature.
2. Incorporate the user's profile like timeline, professional background, Location, and concerns into the recommendations.
3. Clearly connect insights from the abstracts to the user's specific context and goals.
4. Make sure to output in points (1,2,3..) without inserting any **.
5. End your response with a confidence score (in percentage) and a short explanation for that score.

Here are the 5 research abstracts:
1. {lit1}
2. {lit2}
3. {lit3}
4. {lit4}
5. {lit5}

Context: {context}
Question: {question}

Based on the above abstracts, write the answer in points. Make sure to take into account all the information in the context like profession, timeline, etc. Do not include subpoints.

H.2 Specificity

For specificity evaluation, we provide claim, retrieved documents from the knowledge base and specific details to verify.

Prompt for Specificity Evaluation

You are a strict evaluator of specificity and factuality.
Given for each claim:
- A factual claim
- A list of evidence passages from a trusted source
- A set of specific details extracted from the claim (hazard type, location, timeline, intensity)
Your task is to evaluate the claim using ONLY the provided evidence.

LABEL DEFINITIONS

For each specific detail (hazard, location, timeline, intensity), use EXACTLY one of the following labels:
- "yes":
The detail is explicitly mentioned in the claim AND it matches the same specific detail discussed in the knowledge source.
- "no":
The detail is explicitly mentioned in the claim BUT the knowledge source does NOT provide sufficient information to verify it. (This includes cases where the evidence contradicts the claim or does not confirm it.)
- "N/A":
The detail is NOT mentioned in the claim at all.
For location matching, agreement at the STATE level is sufficient; an exact county or city match is not required. Do NOT infer or assume any facts beyond the evidence. Lack of verification MUST be labeled as "no" (not "N/A").

EVALUATION STEPS

Your task is to:
1. Determine whether the **claim is factually true, false, or partially correct**, using ONLY the evidence.
2. For each of the 4 specific details (hazard, location, timeline, intensity):
- Assign "yes", "no", or "N/A" based on the rules above.
- Provide a brief factual explanation for your decision.
3. Justify your overall factuality decision concisely and objectively.
4. If the claim is "true", cite the exact evidence passage(s) that support it.
5. If the claim is "false" or "partially correct", explain precisely which details are unsupported or incorrect.

OUTPUT FORMAT

Return your answer as a SINGLE JSON object in the following format (with no markdown, no extra text, and no explanations outside the JSON):
{{ "claim": "<Claim>",
"hazard": "yes" | "no" | "N/A",
"hazard_reasoning": "<Explain whether hazard mentioned in the claim is explicitly supported>",
"location": "yes" | "no" | "N/A",
"location_reasoning": "<Explain whether location mentioned in the claim is supported>",
"timeline": "yes" | "no" | "N/A",
"timeline_reasoning": "<Explain whether timeline like date and range of years mentioned in the claim is supported>",
"intensity": "yes" | "no" | "N/A",
"intensity_reasoning": "<Explain whether intensity mentioned in the claim is supported>"
}}

INPUTS

Claim: {claim}
Specific Details to Check: {specific_info}
Evidence Passages: {knowledge}

H.3 Answer Relevance

We use LLM for masking important specific details for answer relevance evaluation.

Prompt for Masking

You are a semantic masker. Given the following answer, replace:

- hazard types with [HAZARD]
- profession-related terms with [PROFESSION]
- concern with [CONCERN] (e.g., critical vulnerabilities, maintenance strategies, modernization measures, maintenance strategies, projected impact, design standards, cascading impacts etc.)
- infrastructure with [INFRASTRUCTURE] (e.g., "highway network", "bridge system", "public transit system", "railway infrastructure", "airport facilities", "port facilities", "freight terminals", "traffic control systems", "water treatment plant", "wastewater system", "dam infrastructure", "stormwater system", "coastal protection", "water distribution network", "electrical grid", "power distribution network", "EV charging network", "renewable energy infrastructure", "energy storage facilities", "power transmission lines", "substations", "public buildings", "critical facilities", "commercial structures", etc.)

Keep the structure natural and readable.
Answer: {answer}

H.4 Robustness

For robustness, we use the following paraphrasing and perturbation prompts.

Prompt for Paraphrasing

You are a query paraphrase specialist. Create a variation of the original question by SUBSTITUTING VOCABULARY. Replace keywords with synonyms, use more formal/informal language. Preserve the core meaning and information need. Output ONLY the paraphrased question with no explanations.
Original question:{question}

Prompt for Perturbation

You are a query perturber specialist. Create a variation of the original question by SUBSTITUTING hazard types, and location Output ONLY the perturbed question with no explanations.
Original question:{question}