WXImpactBench: A Disruptive Weather Impact Understanding Benchmark for Evaluating Large Language Models

Yongan Yu[1], Qingchen Hu[1], Xianda Du[2], Jiayin Wang[3], Fengran Mo[4]*, Renée Sieber[1]*

[1]McGill University, [2]University of Waterloo, [3]Tsinghua University, [4]University of Montreal
yongan.yu@mail.mcgill.ca, fengran.mo@umontreal.ca, renee.sieber@mcgill.ca

Abstract

Climate change adaptation requires the understanding of disruptive weather impacts on society, where large language models (LLMs) might be applicable. However, their effectiveness is under-explored due to the difficulty of high-quality corpus collection and the lack of available benchmarks. The climate-related events stored in regional newspapers record how communities adapted and recovered from disasters. However, the processing of the original corpus is non-trivial. In this study, we first develop a disruptive weather impact dataset with a four-stage well-crafted construction pipeline. Then, we propose WXIMPACTBENCH, the first benchmark for evaluating the capacity of LLMs on disruptive weather impacts. The benchmark involves two evaluation tasks, multi-label classification and ranking-based question answering. Extensive experiments on evaluating a set of LLMs provide first-hand analysis of the challenges in developing the understanding of disruptive weather impact and climate change adaptation systems. The constructed dataset and the code for the evaluation framework are available to help society protect against disaster vulnerabilities.

1 Introduction

Climate change adaptation (Karl and Easterling, 1999), referring to actions that help reduce societal vulnerability to climate change, demands a sophisticated understanding of the disruptive weather impacts on society (e.g., the perspective of economy and policy) (Carleton and Hsiang, 2016). Societal reactions to past disruptive weather events are stored in reliable historical sources (Cerveny et al., 2007). Among them, historical newspapers provide irreplaceable information, recording not just meteorological conditions (Gregory and Williams, 1981; Brunet and Jones, 2011), but crucially, how societies adapted and recovered from disasters (Norris et al., 2008; Handmer et al., 2012). In addition, historical newspapers usually report regional disruptive weather impacts with local experiences, which is valuable to understanding long-term climate patterns and social effects (Ogilvie, 2010) but are often absent in official meteorological records (Batlló et al., 2024).

Climate-Related Terms	
Blizzard	**Flood**
Meteorological Meanings	
By *"severe nowstorm" sense*: "five American cowboys lost their lives in the ***blizzard*** which raged in his section during the first fifteen days of November...."	**By *"overflowing water body" sense*:** "An elderly woman drowned in ***flood*** waters, and 30,000 people in Outback areas remained isolated as heavy rain continued..."
Alternative: Ambiguous or Metaphorical Usage	
As sports team name: "one minute remaining in regulation play to give Montreal a 1-1 draw with the Toronto ***Blizzard*** in the lone Canadian Soccer League game..."	**Metaphorical sense:** "People ***flood*** the venue to watch the Jockey Club race..."

Figure 1: Climate-related polysemy examples in different narratives.

Understanding complex patterns in disruptive weather events is important for society with forecasts, societal responses, and public policy (Pielke Jr and Carbone, 2002). The challenge of identifying impacts and responses often lies in climate-related text processing, which contains period-specific narratives and domain-specific linguistic phenomena. For example, disambiguation and taxonomy polysemy can occur in newspaper articles, where climate-related terms frequently appear in diverse linguistic contexts beyond their meteorological meanings. Figure 1 shows that the term "blizzard" can refer to a severe snowstorm or the name of the sports team (e.g., "Toronto Blizzard"). Similarly, the term "flood" can describe an overflowing body of water or can be used metaphorically (e.g., "flood the venue"). This polysemy occurs commonly in newspapers and thus requires the system to distinguish the literal weather-related meanings and alternate usages by improving the

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 4016–4035

climate-related semantic understanding (Nazeer et al., 2024). Another challenge lies in extracting information from the original paper content. Although it is commonly achieved by optical character recognition (OCR) (Thomas et al., 2024), errors remain due to mixed content formats, and complex narrative structures (Nazeer et al., 2024). These errors can negatively affect the extracted text for disruptive weather impact analysis, which renders the texts difficult to serve as a high-quality corpus.

Existing studies on climate-related language processing focus on extracting climate patterns (Alaparthi and Mishra, 2020), wildfire resilience (Xie et al., 2024) and analyzing extreme weather events (Mallick et al., 2024a). Intuitively, LLMs (Törnberg, 2023; Mao et al., 2023; Yang et al., 2024; Mo et al., 2024a) offer a powerful alternative for understanding disruptive weather impacts. However, their effectiveness is unexplored (Boros et al., 2024; Yuan et al., 2025) due to the lack of a corresponding benchmark. The resources used in previous studies cannot comprehensively evaluate the ability of LLMs for weather impacts. This is because i) compared to informative reports in newspapers, previously used meteorological records do not contain long-term and detailed regional information (Pevtsov et al., 2019); and ii) the previous meteorological records are easily obtained and have been available for a long while. Thus they might be already included in the pre-training of LLMs and should not be included in benchmark build-up to avoid potential bias (Ferrara, 2023). To develop a system that assesses the impact of disruptive weather on society, the first step is to establish a domain benchmark for the evaluation protocol.

In this study, we design a four-stage data construction pipeline that begins with a disruptive weather impact dataset in which we correct OCR errors in digitalized newspaper article extraction. We extract a sample of articles from two time periods, which cover linguistic and social changes across different eras and increase linguistic complexity due to the different descriptions of weather events in different times (Campbell, 2013). Historical newspapers often employed more descriptive and elaborate narratives compared to modern reporting styles (Bingham, 2010). These narratives frequently included outdated terminology, spelling variations, and evolving writing conventions (Campbell, 2013). The articles are selected by topic modeling, including six impact categories (infrastructural, political, financial, ecological, agricultural, and human health), which are informed by previous studies (Imran et al., 2016a) and align with modern disaster impact assessment frameworks (Silva et al., 2022).

With our constructed dataset, we develop a benchmark, WXIMPACTBENCH, to investigate the capacity of LLMs to understand disruptive weather impacts with two tasks: i) multi-label classification and ii) ranking-based question-answering. The multi-label classification task employs the previous six impact categories as labels for each article whose ground-truth is annotated by human labor. The question and the candidate pools for the ranking-based question-answering task are constructed based on the context and annotation of the multi-label classification task. This can facilitate any future development of retrieval-augmented generation (RAG) systems in the climate-related domain (Zhao et al., 2024; Mo et al., 2024b; Huang et al., 2024; de Rijke et al., 2025). Extensive experiments on evaluating a set of off-the-shelf LLMs provide first-hand analysis of their capacity to understand disruptive weather impacts and reveal the challenges in developing climate change adaptation systems to help society protect against vulnerabilities from disasters.

Our contributions are summarized as follows: (1) We construct a high-quality disruptive weather impact dataset from digitalized newspaper articles in the climate-related domain with a four-stage pipeline. (2) We propose WXIMPACTBENCH with two typical tasks for evaluating the capacity of LLMs on disruptive weather impact understanding, which is the first benchmark to facilitate the development of such domain-specific systems. The constructed dataset and the evaluation framework code are available at our Github repository[1]. (3) We conduct extensive experiments on benchmarking a set of LLMs, providing first-hand analysis of challenges in disruptive weather impact understanding and climate change adaptation.

2 Related Work

2.1 Climate Impact Analysis and Database

Climate impact analysis (Thulke et al., 2024) aims to help society make correct decisions about climate-related challenges affecting communities, e.g., understanding the weather impacts on society. Existing studies aim to validate the quality of historical weather data (Sieber et al., 2022) or

[1] https://github.com/Michaelyya/WXImpactBench

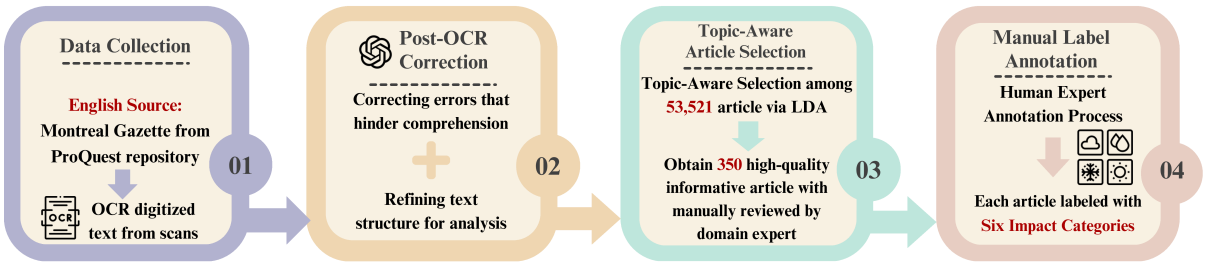


Figure 2: **Data Construction Pipeline** consists of four main stages: (1) Corpus collection from historical newspapers across two periods. (2) Post-OCR correction for high-quality extraction. (3) Article selection with defined categorization using LDA topic modeling and expert curation. (4) Annotation framework conducted by domain experts with a six-category impact classification scheme for understanding disruptive weather impacts.

extract climate patterns via name entities recognition tasks (Mallick et al., 2024b; Xie et al., 2024). Their used corpus is sourced from structured climate science materials, however, usually with daily loss (Batlló et al., 2024), due to the deterioration of storage media (paper, microfiche/microfilm, magnetic tape) (Pevtsov et al., 2019). Compared to structured climate-related scientific databases, historical newspapers can offer a better alternative due to their rich climate records (Vargas-Solar et al., 2021), although they remain largely untapped (Krishnan and Anoop, 2023). The scarcity of high-quality climate-related and nuanced textual data results in the lack of standard benchmarks, which limits understanding of weather impacts. We address these issues in this paper.

2.2 Climate Text Processing and Benchmark

Extracting and processing historical climate articles in newspapers is challenging due to their non-digital formats, such as scanned images or physical archives. OCR enables their conversion into machine-readable text (Baird, 2004), facilitating large-scale digitization, retrieval, and analysis. Like digital libraries (Singh et al., 2012), OCR enhances accessibility, supporting research on climate trends and societal responses. Although neural OCR correction models (Drobac and Lindén, 2020) improve the quality of the extracted text, the degraded print quality, inconsistent terminology, and irregular column layouts (Binmakhashen and Mahmoud, 2019) cause potential errors, which negatively impact the text understanding and the usage for designing downstream tasks (Bingham, 2010; Spathis and Kawsar, 2024; Wang et al., 2024).

Thus, the lack of high-quality resources constrains the development of comprehensive benchmarks for weather impacts. Li et al. (2024) introduce CLLMate, a multimodal benchmark that aligns meteorological data with textual event descriptions for weather event forecasting, though it focuses on prediction rather than historical impact understanding. Developing a benchmark for understanding weather impacts is important, although the fragmented, incomplete, or dispersed disparate sources of weather events increase the difficulty of annotation (Lamb, 2002; Campbell, 2013). Although previous studies (Rasp et al., 2020, 2024) attempt to develop evaluation frameworks for physics-based weather forecasting models, they focus on data-driven weather modeling rather than weather impact understanding. In this paper, we propose the first disruptive weather impact benchmark to fill in this blanks.

3 WXImpactBench: Disruptive Weather Impact Benchmark

Our WXIMPACTBENCH benchmark aims to evaluate to what extent existing LLMs can understand disruptive weather impacts, which also shows the evolution of vulnerability and resilience strategies from society across various periods. It involves two main stages: i) dataset construction; and ii) task definition and evaluation.

3.1 Dataset Construction

The construction of the dataset aims to obtain high-quality text samples. The pipeline overview is presented in Figure 2, which consists of four stages: data collection, post-OCR correction, topic-aware article selection, and manual label annotation.

Data Collection. The data is obtained through collaboration with a proprietary archive institution covering two temporal periods. The original data stored as digitalized text is obtained through OCR (Cheriet et al., 2007), which contains

substantial noise due to historical newspaper layouts, including uneven printing, varying font styles (Sulaiman et al., 2019), complex multicolumn structures (Binmakhashen and Mahmoud, 2019), and overlapping text elements (e.g., advertisements) (Verhoef et al., 2015). Thus, post-OCR correction is necessary to ensure the corpus is high-quality (Chiron et al., 2017; Traub et al., 2015). Our choice of newspapers as the primary data source stems from their status as the most consistent and reliable documentation of social impacts from extreme weather events (Bingham, 2010). After the initial selection, we chose to examine a historical and a modern period, where the historical one was chosen as it had the greatest density of corresponding meteorological information, while the modern one is defined by the Intergovernmental Panel on Climate Change (Seneviratne et al., 2021).

Post-OCR Correction. The goal of post-OCR correction is to correct errors that could significantly impact human comprehension or downstream task analysis, e.g., correct the inaccurate works split and remove unnecessary characters (O'Hara, 2013). For example, given the source image of a digital newspaper article (An example is provided in Appendix A.1), the text obtained by initial OCR and post-OCR correct is shown in Figure 3. The post-OCR correction is achieved by deploying GPT-4o with customized prompts (Zhang et al., 2024). The correction quality is evaluated using BLEU and ROUGE metrics, achieving high consistency scores compared to human annotations, see Appendix A.3 for detailed evaluation results. Our specific prompts for correction are provided in Appendix A.2.

Topic-Aware Article Selection. After the post-OCR correction, we obtain 53,521 weather event-related articles. We aim to obtain informative samples across historical and modern periods based on weather categories. This is achieved by conducting topic modeling on the article collection, where we categorize them via Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to obtain the topic words - representing the primary weather event categories. The details of the categories are provided in Appendix B.1. Selected articles with informative weather content are manually reviewed by three domain experts, which result in 350 high-quality samples. This process ensures the selected articles are topic-aware, which is highly related to specific disruptive weather events.

Initial OCR-Digitized Text

' ', 't 1', ' v MK ' A 0. P. R. 2NGIINE IN THE DIT0H. ' I ', "Th'- snow storm, which was a LL day on Mondy hover' ing overhead, began to set in at dusk, and it gradually increased in Severity, COntInuIng until abt six o-"clock yesterday morning. The storm was the w0rst for many wrintrs, and to find a preedent for it severity IT "1 is said we h' ve to go back to the eventful year ...

After Post-OCR Correction

A C. P. R. ENGINE IN THE DITCH. The snow storm, which was all day on Monday hovering overhead, began to set in about dusk, and it gradually increased in severity, continuing until about six o'clock yesterday morning. The storm was the worst for many winters, and to find a precedent for it in severity it is said we have to go back to the eventful year ...

Figure 3: Example of the text obtained from initial OCR and after our post-OCR correction.

Manual Label Annotation. Having selected article samples, the next step is to assign the annotation for each of them based on the label space. Six vulnerability-related disruptive weather impacts are defined as the labeling categories, including Infrastructural, Political, Financial, Ecological, Agricultural, and Human Health. Annotation is conducted by three domain annotators following our guidelines (provided in Appendix B.2.2). According to the guidelines, the annotators should assign binary labels to indicate the presence or absence of direct descriptions of specific impacts within each article. Unlike previous study (Imran et al., 2016a), however, each sample might correspond to more than one impact.

3.2 Task Definition and Evaluation

After finalizing the data construction, we design the evaluation framework for our benchmark WXIMPACTBENCH. The overview is shown in Figure 4, which contains two tasks, multi-label classification and ranking-based question-answering, to evaluate the capacity of LLMs to understand disruptive weather impacts.

3.2.1 Multi-Label Text Classification

With the annotated weather impact category for each selected article, the intuitive evaluation task is multi-label classification, which aims to test the ability of LLMs to distinguish the disruptive weather impact for each given article. The previous classification tasks in disaster-related natural language processing (Purohit et al., 2013; Imran et al., 2016b) usually focus on modern crisis communica-

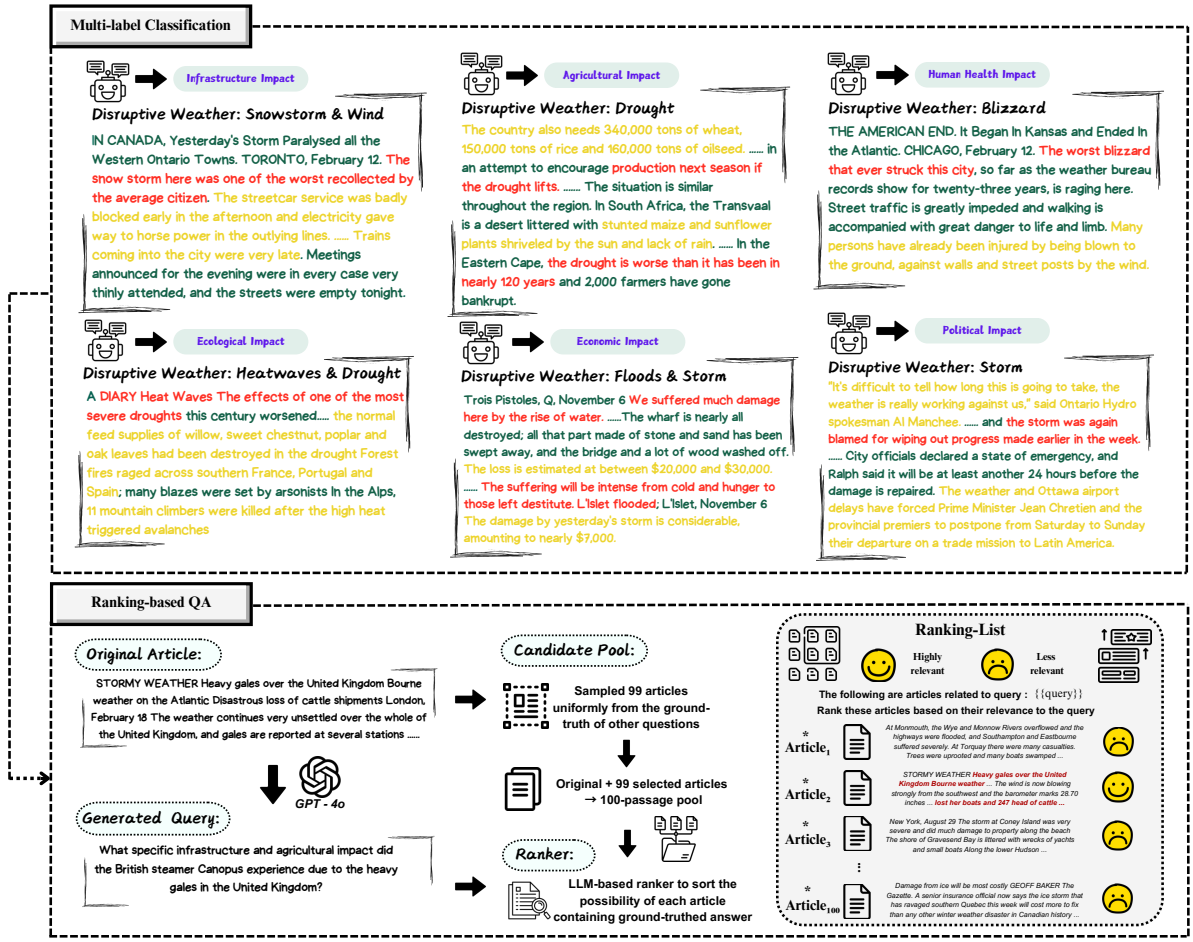


Figure 4: The overview of the benchmark framework with two tasks. Six disruptive weather impacts are used as labeling space in the classification task, where the Red texts represent **disruptive weather events** *(e.g., snowstorm, drought, and blizzard)*, Yellow texts highlight **impact descriptions** *(e.g., damage assessments, resource needs)*, and Green texts refer to **narrative descriptions** *(e.g., geographical locations)*.

tions with structured text. Different from them, our constructed samples require the models to understand the linguistic shifts between historical and modern texts and address inconsistent styles of narratives across various periods. Specifically, given an article sample x_t corresponding to six ground-truth impacts $\mathcal{Y}_t = \{y_t^i\}_{i=1}^6$ with binary labels $y_t^i \in \{0, 1\}$, the evaluated model $\mathcal{M}$ is required to maximize the probability of the predicated impact $\hat{\mathcal{Y}}_t = \{\hat{y}_t^i\}_{i=1}^6$ towards ground-truth. The objective function $\mathcal{L}$ for the given sample x_t of multi-label classification task is formulated as

$$\mathcal{L}(\hat{\mathcal{Y}}_t, \mathcal{Y}_t) = -\sum_{i=1}^{6} y_i \log \hat{y}_i, \quad \hat{y}_i = \mathcal{M}(x_t)$$

3.2.2 Ranking-based Question Answering

Question-answering (QA) requires the LLMs to reply to the given question based on their parametric knowledge. We formulate the ranking-based QA task by prompting the models to identify the likelihood of each article containing the correct answer from a candidate pool. This setting could also facilitate RAG systems development in the domain (Mao et al., 2024; Mo et al., 2024c, 2025), where we left the answer span extraction/generation for future studies.

To construct an evaluation protocol, the first step is to obtain suitable question pairs with each annotated samples in the multi-label classification task, since the question set is unavailable. Thus, we generate pseudo questions q_t for each article $(x_t, \mathcal{Y}_t)$ based on its annotated category via a generative LLM $\mathcal{G}$ which is formulated as $q_t = \mathcal{G}(x_t, \mathcal{Y}_t)$. The annotated categories $\mathcal{Y}_t$, which are the societal impacts brought by the disruptive weather event, will become part of the prompt to ensure the generated question targets one of the specific impact categories (see Figure 4).

As a result, we have QA pair (q_t, x_t) for each sample. The next step is to construct the candidate pool for ranking. The size of the pool $\mathcal{X}_t$ for

each question q_t is 100, which contains the ground-truth x_t and other 99 counterexamples $\mathcal{X}_t^-$ that are randomly sampled from the ground-truth of other questions. With the constructed QA pairs and corresponding candidate pools, the evaluated model $\mathcal{M}$ is required to rank the ground-truth based on the relevant scores produced by a matching function $\mathcal{S}$. The task objective can be formulated as

$$\text{RankingList } \phi(q_t) \leftarrow \underset{\mathcal{S}(q_t, x_t)}{\arg\max}\ \mathcal{M}(q_t, \mathcal{X}_t, \mathcal{S})$$

where $\{x_t, \mathcal{X}_t^-\} \subseteq \mathcal{X}_t$ and the output ranking list $\phi(q_t)$ is evaluated by ranking metrics.

4 Experimental Setup

4.1 Evaluation Metrics and Settings

For multi-label classification task, we use F1-score, accuracy, and row-wise accuracy as evaluation metrics. The evaluation via F1-score and accuracy are averaged across the six impact categories, historical and modern articles, and the effect of different context lengths. Compared to the common F1-score and accuracy, the row-wise accuracy is a strict metric that requires more accurate output as the model should correctly classify all six impact labels for a given article, defined as

$$\text{Row-wise Acc.} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{6}\mathcal{I}\left(\hat{y}_i^j = y_i^j\right)$$

where N is the number of samples, $\hat{y}_i^j$ denotes the predicted label for the j-th category in the i-th sample, y_i^j is the corresponding ground-truth label, and $\mathcal{I}(\cdot)$ is the indicator function. The evaluation goal is to investigate the models' ability to distinguish various disruptive weather impacts under different settings, e.g., different periods.For the long-context impact evaluation, we create an alternate version (mixed context), whose sample length is split into segments with approximately 250 tokens from the original one (long-context version) following (Levy et al., 2024). Note that annotations for these smaller chunks are performed independently by the same domain experts rather than automatically inherited from the original articles. This independent annotation process naturally results in some chunks containing no weather impact labels, which serve as valuable negative examples in our evaluation. These negative samples allow us to assess models' understanding to correctly identify passages without weather impacts, particularly reflected in our row-wise accuracy metric. Eventually, we contain 350 and 1,386 samples for the original and mixed context version datasets, respectively. The detailed statistics of the datasets are provided in Appendix D.

For the ranking-based QA task, we deploy the standard metric that emphasizes the accuracy of top positions for evaluation, including Hit@1, nDCG@5, Recall@5, and MRR.

4.2 Evaluated Models

We evaluate a set of off-the-shelf LLMs on WXIMPACTBENCH. For the multi-label text classification task, we include seven open-source models: DEEPSEEK-V3-671B (DeepSeek-AI, 2024), LLAMA-3.1-8B-INSTRUCT (Llama, 2024), Mistral-7B-Instruct (Jiang et al., 2023), MIXTRAL-8X7B-INSTRUCT (Jiang et al., 2024), MISTRAL-24B-INSTRUCT (Jiang et al., 2024), GEMMA-2-9B-IT (GemmaTeam, 2024), QWEN2.5-7B-INSTRUCT, and QWEN2.5-14B-INSTRUCT (Qwen2.5, 2025); and three closed-source models: GPT-3.5-TURBO, GPT-4 (OpenAI, 2024a), and GPT-4O (OpenAI, 2024b). For the ranking-based QA task, we evaluate GPT-3.5-TURBO, QWEN2.5-7B-INSTRUCT, QWEN2.5-14B-INSTRUCT, MISTRAL-7B-INSTRUCT, and LLAMA-3.1-8B-INSTRUCT. The relatively smaller models (with 7B size) ensure the latency requirements (Sun et al., 2023). The used models for the two tasks cover different sizes and support the input length of at least 8k tokens. The version details are provided in Appendix E.

4.3 Implementation Details

Multi-label Classification. The multi-label classification is conducted on each evaluated LLM by the same prompt provided in Appendix C.2. Different from traditional methods that decompose multi-label text classification into multiple binary classification tasks (Boutell et al., 2004; Liu et al., 2017), we simultaneously identify all relevant disruptive weather impacts for each input by calling the LLM once. The example of in-context learning in the one-shot setting is handcrafted with a complex sample detailing multiple impacts.

Ranking-based Question-Answering. We employ GPT-4O for pseudo question generation with default hyper-parameters. For ranking evaluation, we adopt the sliding window mechanism within

	Infrastructure	Political	Financial	Ecological	Agricultural	Human Health	Average
Positive Cases	168 (326)	61 (101)	98 (134)	54 (71)	80 (100)	117 (178)	-
Zero-Shot Performance							
GPT-4o	80.94 ↑ 0.39	**58.46** ↑ 3.70	**65.82** ↓ 0.23	**46.81** ↑ 2.51	**70.33** ↑ 1.92	**73.23** ↑ 1.12	**65.93** ↑ 1.07
GPT-4	74.87 ↑ 2.13	49.38 ↑ 2.47	55.70 ↑ 1.18	37.84 ↑ 4.71	60.00 ↑ 3.83	62.96 ↑ 3.71	56.79 ↑ 2.12
GPT-3.5-TURBO	77.59 ↑ 3.73	41.60 ↑ 7.73	42.39 ↑ 7.16	36.52 ↓ 0.01	55.63 ↑ 4.49	47.29 ↑ 10.91	50.17 ↑ 5.12
DEEPSEEK-V3-671B	**81.87** ↑ 0.80	44.44 ↑ 12.40	60.91 ↑ 3.91	36.00 ↓ 0.60	61.74 ↑ 4.34	65.20 ↑ 0.20	58.03 ↑ 3.07
MISTRAL-24B-IT	79.12 ↓ 0.17	47.18 ↑ 6.91	59.64 ↑ 2.04	44.90 ↑ 10.75	67.74 ↑ 1.07	66.88 ↑ 1.30	60.91 ↑ 1.57
MIXTRAL-8X7B-IT	72.31 ↑ 3.01	39.29 ↑ 6.86	57.02 ↑ 0.33	36.59 ↑ 5.23	44.44 ↑ 14.23	50.00 ↑ 9.69	49.94 ↑ 4.84
MISTRAL-7B-IT	74.27 ↑ 4.27	36.63 ↑ 6.63	45.56 ↑ 0.56	39.19 ↑ 5.19	55.30 ↑ 5.30	61.61 ↑ 11.61	52.76 ↑ 1.98
QWEN2.5-14B-IT	76.21↓ 1.45	41.45 ↑ 0.34	45.07 ↑ 5.32	41.62 ↓ 2.15	52.99 ↑ 5.01	63.08 ↓ 1.64	53.74 ↓ 1.36
QWEN2.5-7B-IT	70.52 ↑ 4.75	34.29 ↑ 2.82	43.43 ↑ 0.83	41.06 ↓ 6.51	40.26 ↑ 3.96	38.19 ↑ 7.34	44.63 ↑ 1.88
GEMMA-2-9B-IT	77.42 ↑ 1.52	43.33 ↑ 2.59	54.60 ↓0.33	42.16 ↑ 1.73	55.60 ↑ 4.10	61.82↑ 0.87	55.82↑ 1.31
LLAMA-3.1-8B-IT	70.13 ↑ 8.82	40.47 ↑ 1.51	55.29 ↓ 2.08	33.90 ↑ 4.99	55.49 ↑ 5.05	50.68 ↑ 11.72	50.66 ↑ 4.91
Average	75.93 ↑ 2.93	43.32 ↑ 5.36	53.22 ↑ 1.34	43.33 ↑ 2.80	56.32 ↑ 4.83	58.36 ↑ 5.85	54.48 ↑ 2.38
In-Context Learning (One-shot) Performance							
GPT-4o	81.25 ↑ 0.29	**59.54** ↑ 2.71	**63.64** ↑ 1.18	**50.00** ↓ 0.02	**71.43** ↑ 1.39	**72.94** ↑ 2.80	**66.93** ↑ 1.02
GPT-4	72.63 ↑ 3.54	40.00 ↑ 5.92	55.15 ↑ 0.55	32.38 ↑ 7.30	61.29 ↑ 2.92	60.22 ↑ 5.06	54.95↑ 3.25
GPT-3.5-TURBO	76.88 ↑ 2.29	38.93 ↑ 9.72	48.50 ↑ 0.95	40.00 ↑ 0.02	57.30 ↓ 0.02	60.98 ↓ 0.02	54.08↑ 2.26
DEEPSEEK-V3-671B	**81.62** ↑ 1.32	49.48 ↑ 7.40	63.37 ↑ 2.10	43.55 ↑ 1.21	62.82 ↑ 5.04	67.78 ↑ 3.07	62.90 ↑ 2.36
MISTRAL-24B-IT	78.38 ↑ 0.58	43.48 ↑ 12.36	56.99 ↑ 1.05	35.09 ↑ 6.39	61.45 ↑ 4.50	65.28 ↓ 0.08	57.41↑ 2.06
MIXTRAL-8X7B-IT	68.31 ↑ 4.47	12.50 ↑ 24.14	42.00 ↑ 8.43	26.45 ↑ 7.74	36.80 ↑ 10.26	46.46 ↑ 14.82	40.53↑ 8.63
MISTRAL-7B-IT	73.31 ↑ 3.31	20.74 ↑ 6.74	45.33 ↑ 5.33	31.94 ↑ 1.94	52.77 ↑ 2.77	54.87 ↑ 4.87	47.43↑ 1.57
QWEN2.5-14B-IT	78.10 ↑ 0.05	43.36 ↑ 1.21	48.42 ↑ 6.13	43.65 ↑ 0.30	62.18 ↑ 4.49	63.60 ↑ 3.78	54.95 ↑ 1.71
QWEN2.5-7B-IT	71.04 ↑ 2.70	31.46 ↓ 4.19	48.80 ↑ 0.69	37.68 ↓ 0.09	47.54 ↓ 14.21	45.85 ↑ 8.51	47.40 ↑ 0.75
GEMMA-2-9B-IT	74.24 ↑ 1.54	31.79 ↑ 7.36	51.76 ↑ 0.91	34.52 ↑ 0.39	48.13 ↑ 7.87	63.76 ↑ 0.57	48.20↑ 1.63
LLAMA-3.1-8B-IT	71.88 ↑ 4.73	34.92 ↑ 8.01	49.50 ↑ 3.95	40.30 ↑ 1.08	52.69 ↑ 3.91	54.85 ↑ 5.48	51.33 ↑ 4.45
Average	74.27 ↑ 2.48	35.07 ↑ 8.02	52.38 ↑ 2.51	37.32 ↑ 2.60	55.31 ↑ 2.44	58.95 ↑ 4.51	56.63↑ 2.70

Table 1: F1-scores results of zero-shot and one-shot evaluation categorized on six impacts and two context length settings. The number in parentheses refers to the improvement with ↑ or degradation with ↓ of the evaluation on the mixed-context dataset (1,386 entries) compared to the original dataset (350 entries). The number in the "Positive Cases" row denotes the positive label in each impact categorization corresponding to two context-length versions. **Bold** and underline indicate the best and the second-best performance.

LLM-based ranker implementation following the state-of-the-art study (Sun et al., 2023) to reduce the potential negative effect of noisy long contexts. Specifically, each article in the candidate pool was segmented into three chunks, and then the initial ranking was performed independently within each chunk. The used prompts for both stages are provided in Appendix C.3.

To ensure stable results, following previous studies (Chen et al., 2023), all LLMs were evaluated with the temperature set to 0, and the reported performance is the average value of running the experiments three times.

5 Experiments

5.1 Results of Multi-label Classification

Table 1 and Table 2 show the performance of the evaluated LLMs on WXIMPACTBENCH for the settings of categorized by six societal impacts with different context lengths, overall row-wise evaluation, and divided into two periods, respectively. The additional experimental results are provided in Appendix G. We have the observations as follows.

LLMs struggle to understand disruptive weather impacts. Table 1 shows that the F1-score for multi-label classification remains consistently low across models, especially among the political and ecological categories. The financial, agricultural, and human health impacts categories perform slightly better but still exhibit suboptimal results at 55%. The low performance might be attributed to the challenges in these categories with abstract and context-dependent narratives. Different from the infrastructure category (with the high-

Model	Zero-Shot	One-Shot
GPT-4o	32.28 ↑ 0.29	**31.99** ↓ 0.85
GPT-4	22.19 ↑ 0.38	20.46 ↑ 0.11
GPT-3.5-TURBO	21.61 ↓ 0.18	12.39 ↑ 6.18
DEEPSEEK-V3-671B	**34.01** ↓ 1.72	**31.99** ↓ 0.28
MISTRAL-24B-IT	19.88 ↓ 1.02	25.65 ↓ 1.08
MIXTRAL-8X7B-IT	25.07 ↓ 0.50	19.88 ↑ 0.12
MISTRAL-7B-IT	4.90 ↓ 0.04	8.93 ↓ 3.50
QWEN2.5-14B-IT	19.02 ↓ 2.45	18.16 ↓ 0.73
QWEN2.5-7B-IT	27.38 ↓ 6.52	25.65 ↓ 1.36
GEMMA-2-9B-IT	15.56 ↑ 0.44	9.51 ↓ 1.51
LLAMA-3.1-8B-IT	12.68 ↑ 2.18	15.56 ↓ 1.85
Average	21.96 ↓ 0.57	20.80 ↓ 0.10

Table 2: Row-wise accuracy performance across different models and prompting strategies. The used notions are the same as Table 1.

est performance), which describes the impacts of weather events explicitly, e.g., "bridges destroyed" and "power outages", the descriptions in other categories are usually more abstract. For example, the financial damage could be embedded within discussions of damaged infrastructure or agricultural setbacks, which makes it more difficult for models to distinguish them as standalone impacts.

Table 2 shows row-wise performance, in which the model must identify the given sample correctly for each involved category, the performance of classification drops dramatically due to the more precise requirement. Thus, a sophisticated model is expected to understand the complex societal effects of historical narratives via reasoning (Wei et al., 2022; Zhang et al., 2025a,b).

Long-context LLMs not always be strong on long-context de-noising. The results in Table 1 show that, when the original long-context is segmented into smaller chunks, the classification accuracy increases in most cases. These improvements suggest that smaller chunks help models focus on relevant information and thus minimizing distraction from extraneous content. Even the used models are claimed with long-context capacity, more precise split that reduces potential noise is still effective for context de-noising, which is consistent with previous studies (Sun et al., 2024).

However, we also find that this trend is not observed with the row-wise accuracy evaluation. This is due to the evaluation bias, where the F1-score measures precision and recall per category, and benefits from partial correctness. Row-wise accuracy requires an exact match across all labels. The small chunks might be helpful to improve the classification of one of the categories but not enough to correct all labels. Thus, the helpfulness of long-context de-noising via small chunks is not obvious in overall performance.

In-context Learning offers limited improvement. The in-context learning is achieved by providing one demonstration as the one-shot example for model decision. Compared zero-shot and one-shot performance in Table 1, we find that providing a single example in the prompt offers limited benefits and might decrease the performance in some cases. Such a phenomenon implies that the LLMs lack sufficient knowledge to disambiguate disruptive weather impacts even with enhanced examples for knowledge arousing.

	Historical	**Modern**
Cases	200 (504)	150 (882)
GPT-4o	**70.19** ↑ 1.50	**68.59** ↑ 0.59
GPT-4	65.54 ↑ 0.66	53.81 ↑ 4.50
GPT-3.5-TURBO	57.27 ↑ 4.78	52.21 ↑ 5.92
DEEPSEEK-V3-671B	65.42 ↑ 3.91	64.74 ↑ 2.05
MISTRAL-24B-IT	62.30 ↑ 4.13	63.33 ↓ 1.68
MIXTRAL-8X7B-IT	58.31 ↑ 4.28	50.11 ↑ 6.14
MISTRAL-7B-IT	55.27 ↑ 1.50	49.22 ↑ 2.78
QWEN2.5-14B-IT	57.75 ↑ 2.30	57.75 ↓ 5.48
QWEN2.5-7B-IT	54.29 ↑ 0.66	40.85 ↑ 4.14
GEMMA-2-9B-IT	60.41 ↑ 0.90	52.80 ↑ 1.97
LLAMA-3.1-8B-IT	55.30 ↑ 5.08	50.67 ↑ 4.98
Average	60.19 ↑ 2.70	54.92 ↑ 2.36

Table 3: F1-score performance across historical and modern impact categories in zero-shot setting. The used notations are the same as Table 1.

Model	**Hit@1**	**nDCG@5**	**Recall@5**	**MRR**
GPT-3.5-TURBO	62.09	67.31	71.04	66.90
MISTRAL-7B	6.21	15.86	25.16	14.82
QWEN-2.5-14B	**82.09**	**86.34**	**85.48**	**89.55**
QWEN-2.5-7B	42.69	61.80	75.52	58.04
LLAMA-3.1-8B	64.18	70.85	75.82	69.90
Result with Bias for Reference				
GPT-4o	91.94	95.54	97.91	94.88

Table 4: The performance of ranking-based QA tasks across different models.

These results indicate that our WXIMPACTBENCH is challenging for LLMs to understand disruptive weather impact.

Historical narratives are easier for LLMs to understand. The evaluation of different narratives in terms of historical and modern articles is presented in Table 3. Surprisingly, the evaluated models perform better on the articles recorded in the historical period. The reason might be the structured and formal narrative style used to report disruptive weather events in historical periods, which more explicitly highlights cause-and-effect relationships. The observation is revealed by the earlier studies (e.g., Mauch and Pfister, 2009), where the historical narratives emphasize empirical observations over interpretations, offering a more immediate and naturalistic account of events. Though the modern text might dominate within the pre-trained corpus, the language patterns used in historical narrative styles are easier for language models to identify, and thus perform better on classifying disruptive weather impacts.

Scaling law might hold for disruptive weather impact understanding. As illustrated in Table 1

and Table 2, larger models usually perform better than smaller ones, which is consistent with the scaling law for LLMs (Kaplan et al., 2020). For example, the largest DEEPSEEK-V3-671B obtains the best results and MISTRAL-24B-IT outperforms its 7B version. Although the model size is unavailable in closed-source models, the open-source models with the feasibility of manipulation can be viable alternatives to adaptively work for domain requirements. With proper optimization, the second-best DEEPSEEK-V3-671B for understanding disruptive weather impact might offer performance close to or on par with GPT-4O.

5.2 Results of Ranking-based QA

The performance of each evaluated model for ranking-based QA is reported in Table 4. We find that QWEN-2.5-14B-IT achieves the best performance in this task. LLAMA-3.1-8B is slightly better than GPT-3.5-TURBO and QWEN-2.5-7B-IT, while the MISTRAL-7B-IT cannot address the tasks related to disruptive weather context. Notice that the ranking results would contain bias when the evaluated model is used for question generation (GPT-4O in our cases). This is a common phenomenon (Zhou et al., 2023) and needs to be avoided in benchmarking.

The practical open-retrieval setting, i.e., identifying the relevant articles from a huge database, is left for future studies, which could further facilitate knowledge enhancement in understanding disruptive weather impacts.

6 Conclusion

In this study, we propose a disruptive weather impact understanding benchmark, WXIMPACTBENCH, to address the lack of datasets and evaluation frameworks in climate change adaptation. We first process the corpus from newspaper articles and provide comprehensive instruction for impact annotation with each processed article.

Then, we design multi-label classification and ranking-based QA tasks to evaluate the LLMs in understanding various defined disruptive weather impacts. Extensive experiments on WXIMPACTBENCH reveal that the existing LLMs struggle with understanding disruptive weather impacts across different style narratives and context settings. Our WXIMPACTBENCH enables the community to evaluate the developed systems on disruptive weather impact understanding, which supports the society to learn from and prepare for the impacts of climate change.

Limitation

Although WXIMPACTBENCH provides valuable insights (e.g., exhibit the strengths and weaknesses of various society impact understanding) about evaluating LLMs on disruptive weather, it may have potential biases in underrepresented historical events and linguistic variations. Future work could expand the range of evaluated models, strategies, and designed tasks to further strengthen the evaluations.

Ethical Considerations

Our primary data source is a corpus of historical digitized newspapers, obtained through collaboration with an official organization, which should be anonymous at this moment. This organization preserves the copyright of the newspaper articles and has been granted permission to publish this subset of articles for benchmark build-up to facilitate the research community. Thus, the data is publicly available and thus no potential privacy or content safety concerns. Additionally, topic-aware article selection is conducted by researchers specializing in historical climate analysis to ensure the dataset is not biased on specific time and location. This research contributes to the broader societal goal of understanding historical disruptive weather impacts to help society defend its vulnerabilities from disasters. The interpretation of weather-related disruptions in historical newspapers might be influenced by demographic and contextual factors, which is similar to other text datasets generated through crowd-sourcing with inherent challenges in ensuring that dataset labels are fully representative of diverse societal perspectives (Talat et al., 2022).

Acknowledgment

Our primary data source is a corpus of three digitized newspapers (La Presse, La Patrie and Montreal Gazette), obtained through collaboration with the McGill University Library and Archives and the Bibliothèque nationale du Québec. We would like to thank DRAW McGill for their guidance throughout this project, especially Dr. Victoria Slonosky, whose expertise in historical meteorology was instrumental in accessing the corpus and advising on OCR correction. We are also deeply

grateful for the support provided by OpenAI Grants and RBC Borealis AI.

References

Shivaji Alaparthi and Manit Mishra. 2020. Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. *arXiv preprint arXiv:2007.01127*.

Henry S Baird. 2004. Difficult and urgent open problems in document image analysis for libraries. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pages 25–32. IEEE.

Josep Batlló, Hisashi Hayakawa, Victoria Slonosky, and Richard I Crouthamel. 2024. Preface to the special issue on "old records for new knowledge". *Geoscience Data Journal*.

Adrian Bingham. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–231.

Galal M Binmakhashen and Sabri A Mahmoud. 2019. Document layout analysis: a comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771.

Manola Brunet and Phil Jones. 2011. Data rescue initiatives: bringing historical climate data into the 21st century. *Climate Research*, 47(1-2):29–40.

Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press.

Tamma A Carleton and Solomon M Hsiang. 2016. Social and economic impacts of climate. *Science*, 353(6304):aad9837.

Randall S Cerveny, Jay Lawrimore, Roger Edwards, and Christopher Landsea. 2007. Extreme weather records: Compilation, adjudication, and publication. *Bulletin of the American Meteorological Society*, 88(6):853–860.

shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44502–44523. Curran Associates, Inc.

Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, and Ching Suen. 2007. *Character recognition systems: a guide for students and practitioners*. John Wiley & Sons.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of ocr errors on the use of digital libraries: towards a better access to information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4. IEEE.

Maarten de Rijke, Bart van den Hurk, Flora Salim, Alaa Al Khourdajie, Nan Bai, Renato Calzone, Declan Curran, Getnet Demil, Lesley Frew, Noah Gießing, et al. 2025. Information retrieval for climate impact. *arXiv preprint arXiv:2504.01162*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Senka Drobac and Krister Lindén. 2020. Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition (IJDAR)*, 23(4):279–295.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*.

GemmaTeam. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

KJ Gregory and RF Williams. 1981. Physical geography from the newspaper. *Geography*, pages 42–52.

John Handmer, Yasushi Honda, Zbigniew W Kundzewicz, Nigel Arnell, Gerardo Benito, Jerry Hatfield, Ismail Fadl Mohamed, Pascal Peduzzi, Shaohong Wu, Boris Sherstyukov, et al. 2012. Changes in impacts of climate extremes: human systems and ecosystems. *Managing the risks of extreme events and disasters to advance climate change adaptation special report of the intergovernmental panel on climate change*, pages 231–290.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

M Imran, P Mitra, and C Castillo. 2016a. Twitter as a lifeline: Humanannotated twitter corpora for nlp of crisis-related messages.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016b. Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1638–1643, Portorož, Slovenia. European Language Resources Association (ELRA).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Thomas R Karl and David R Easterling. 1999. Climate extremes: selected review and future research directions. *Climatic change*, 42(1):309–325.

Ajay Krishnan and VS Anoop. 2023. Climatenlp: Analyzing public sentiment towards climate change using natural language processing. *arXiv e-prints*, pages arXiv–2310.

Hubert H Lamb. 2002. *Climate, history and the modern world*. Routledge.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

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

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 115–124, New York, NY, USA. Association for Computing Machinery.

Llama. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tanwi Mallick, Joshua David Bergerson, Duane R Verner, John K Hutchison, Leslie-Anne Levy, and Prasanna Balaprakash. 2024a. Understanding the impact of climate change on critical infrastructure through nlp analysis of scientific literature. *Sustainable and Resilient Infrastructure*, pages 1–18.

Tanwi Mallick, John Murphy, Joshua David Bergerson, Duane R Verner, John K Hutchison, and Leslie-Anne Levy. 2024b. Analyzing regional impacts of climate change using natural language processing techniques. *arXiv preprint arXiv:2401.06817*.

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large language models know your contextual search intent: A prompting framework for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1211–1225.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024. Ragstudio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 725–735.

Christof Mauch and Christian Pfister. 2009. *Natural disasters, cultural responses: case studies toward a global environmental history*. Lexington Books.

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Ranked list truncation for large language model-based re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 141–151.

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024a. Chiq: Contextual history enhancement for improving query rewriting in conversational search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2268.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024b. A survey of conversational search. *arXiv preprint arXiv:2410.15576*.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012.

Fengran Mo, Gao Yifan, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, Bing Yin, and Jiang Meng. 2025. Uniconv: Unifying retrieval and response generation for large language model in conversation. In *Proceedings of the 63st Annual Meeting of the Association for Computational Linguistics: ACL 2025*.

Fengran Mo, Longxiang Zhao, Kaiyu Huang, Yue Dong, Degen Huang, and Jian-Yun Nie. 2024c. How to leverage personal textual knowledge for personalized conversational information retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3954–3958.

Imran Nazeer, Khadija Ghulam Hussain, and Saima Jamshaid. 2024. The evolution of linguistic strategies in digital news discourse: A comparative analysis. *International Journal of Human and Society*, 4(1):917–930.

Fran H Norris, Susan P Stevens, Betty Pfefferbaum, Karen F Wyche, and Rose L Pfefferbaum. 2008. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *American journal of community psychology*, 41:127–150.

Astrid EJ Ogilvie. 2010. Historical climatology, climatic change, and implications for climate science in the twenty-first century. *Climatic Change*, 100(1):33–47.

Laura Turner O'Hara. 2013. Cleaning ocr'd text with regular expressions. *The Programming Historian*.

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

OpenAI. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Alexei Pevtsov, Elizabeth Griffin, Jonathan Grindlay, Stella Kafka, Jennifer Lynn Bartlett, Ilya Usoskin, Kalevi Mursula, Sarah Gibson, Valentin M Pillet, Joan Burkepile, et al. 2019. Historical astronomical data: urgent need for preservation, digitization enabling scientific exploration. *arXiv preprint arXiv:1903.04839*.

Roger Pielke Jr and Richard E Carbone. 2002. Weather impacts, forecasts, and policy: An integrated perspective. *Bulletin of the American Meteorological Society*, 83(3):393–406.

Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, 19(1).

Qwen2.5. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11).

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. 2024. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Preprint*, arXiv:2308.15560.

Sonia I Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, A Di Luca, Subimal Ghosh, Iskhaq Iskandar, James Kossin, Sophie Lewis, et al. 2021. Weather and climate extreme events in a changing climate.

Renée Sieber, Victoria Slonosky, Linden Ashcroft, and Christa Pudmenzky. 2022. Formalizing trust in historical weather data. *Weather, Climate, and Society*, 14(3):993–1007.

Vitor Silva, Svetlana Brzev, Charles Scawthorn, Catalina Yepes, Jamal Dabbeek, and Helen Crowley. 2022. A building classification system for multi-hazard risk assessment. *International Journal of Disaster Risk Science*, 13(2):161–177.

Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. 2012. A survey of ocr applications. *International Journal of Machine Learning and Computing*, 2(3):314.

Dimitris Spathis and Fahim Kawsar. 2024. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *Journal of the American Medical Informatics Association*, 31(9):2151–2158.

Alaa Sulaiman, Khairuddin Omar, and Mohammad F Nasrudin. 2019. Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. *Journal of Imaging*, 5(4):48.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2024. Is chatgpt good at search? investigating large language models as re-ranking agents. *Preprint*, arXiv:2304.09542.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging llms for post-ocr correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 116–121.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

Petter Törnberg. 2023. How to use llms for text analysis. *arXiv preprint arXiv:2307.13106*.

Myriam C Traub, Jacco Van Ossenbruggen, and Lynda Hardman. 2015. Impact analysis of ocr quality on research tasks in digital archives. In *Research and Advanced Technology for Digital Libraries: 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19*, pages 252–263. Springer.

Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, Javier A Espinosa-Oviedo, and Luis M Vilches-Blázquez. 2021. Laclichev: Exploring the history of climate change in latin america within newspapers digital collections. In *European Conference on Advances in Databases and Information Systems*, pages 121–132. Springer.

Jesper Verhoef et al. 2015. The cultural-historical value of and problems with digitized advertisements: Historical newspapers and the portable radio, 1950-1969. *TS: Tijdschrift Voor Tijdschriftstudies*, 38:51–60.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yangxinyu Xie, Bowen Jiang, Tanwi Mallick, Joshua David Bergerson, John K Hutchison, Duane R Verner, Jordan Branham, M Ross Alexander, Robert B Ross, Yan Feng, et al. 2024. Wildfiregpt: Tailored large language model for wildfire analysis. *arXiv preprint arXiv:2402.07877*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Long Yuan, Fengran Mo, Kaiyu Huang, Wenjie Wang, Wangyuxuan Zhai, Xiaoyu Zhu, You Li, Jinan Xu, and Jian-Yun Nie. 2025. Omnigeo: Towards a multimodal large language models for geospatial artificial intelligence. *arXiv preprint arXiv:2503.16326*.

James Zhang, Wouter Haverals, Mary Naydan, and Brian W. Kernighan. 2024. Post-ocr correction with openai's gpt models on challenging english prosody texts. In *DocEng 2024 - Proceedings of the 2024 ACM Symposium on Document Engineering*, DocEng 2024 - Proceedings of the 2024 ACM Symposium on Document Engineering. Association for Computing Machinery, Inc. Publisher Copyright: © 2024 Copyright held by the owner/author(s).; 2024 ACM Symposium on Document Engineering, DocEng 2024 ; Conference date: 20-08-2024 Through 23-08-2024.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025a. Entropy-based exploration conduction for multi-step reasoning. *arXiv preprint arXiv:2503.15848*.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025b. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

Appendix

A Post-OCR Correction

A.1 Example of OCR-Digitized Text

Figure 5 presents an example of OCR-digitized text from the *Illustrated Montreal Gazette*, dated February 18, 1885. This excerpt, titled "Snowstorm_delays," was retrieved from *ProQuest* and illustrates the typical noise and distortions introduced by OCR processing in historical newspaper archives.

A.2 Post-OCR Correction Instruction

To reduce the substantial noise in OCR-digitized text, GPT-4o was used for post-OCR correction to enhance text quality. The specific prompt used for this process is presented in Table 5.

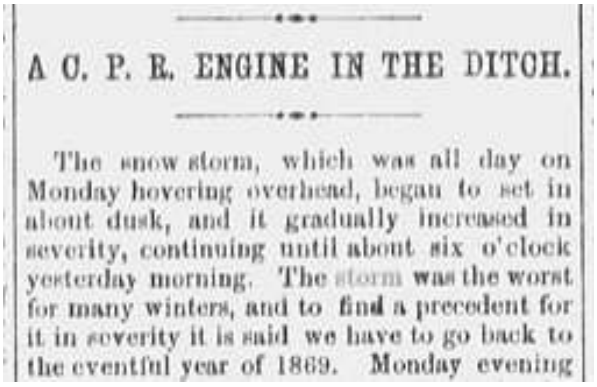
Post-OCR Correction Instruction

You are an expert OCR correction assistant specializing in historical newspaper text. Your task is to:

1. Correct OCR errors while preserving the original text's meaning, structure, and formatting.
2. Accurately retain proper nouns, dates, numbers, and domain-specific terminology.
3. Maintain paragraph breaks, section headers, bylines, and other structural elements.
4. Remove extraneous characters (e.g., unnecessary punctuation, OCR artifacts) without altering the content.
5. Properly reconstruct hyphenated words that were split across lines.
6. Standardize spacing by eliminating extra spaces and ensuring a consistent format.
7. Return the corrected text as a single continuous line, with no newline characters.

NOTE: Do not include explanations, summaries, or additional comments. Only return the corrected text in the specified format.

Table 5: Prompts used for Post-OCR correction.



A C. P. R. ENGINE IN THE DITCH.

The snow storm, which was all day on Monday hovering overhead, began to set in about dusk, and it gradually increased in severity, continuing until about six o'clock yesterday morning. The storm was the worst for many winters, and to find a precedent for it in severity it is said we have to go back to the eventful year of 1869. Monday evening

Figure 5: Example of OCR-digitized text from the *Illustrated Montreal Gazette*, dated February 18, 1885.

A.3 Post-OCR Correction Evaluation

To assess the effectiveness of our post-OCR correction process, we evaluated GPT-4o's output against human-annotated corrections on a randomly selected sample of 50 articles. The results demonstrate the high accuracy of the automated corrections:

Metric	1-gram	2-gram	3-gram / L
BLEU	0.9115	0.8935	0.8773
ROUGE	0.9476	0.9190	0.9438

Table 6: Evaluation metrics comparing GPT-4o OCR corrections to human annotations (50 samples). ROUGE-L is listed under the third column.

The consistently high BLEU and ROUGE scores indicate that GPT-4o's corrections closely align with human-edited versions, validating its effectiveness for improving text quality prior to downstream analysis.

B Annotation Guidelines for Multi-Label Classification

B.1 Definition of Primary Disaster Categories

Using Latent Dirichlet Allocation, the dataset was categorized into 15 primary weather event types. The major weather categories are listed in Table 7:

B.2 Background

In the absence of standardized impact records (e.g., flood-related property damage, injuries due to ice accumulation, power outages, and road closures), we assessed vulnerabilities and resilience based on the consequences of weather events and how they have changed since the 19th century. To do so, we categorized disruptive weather impacts into six primary groups — Infrastructural, Agricultural,

Disaster Type	Definition	Example
Blizzard	Severe snowstorm	*Whiteout conditions*
Cold	Low temperatures	*Frostbite risk*
Deluge	Heavy rainfall	*Flash flooding*
Drought	Prolonged dryness	*Water scarcity*
Flood	Overflowing water	*River flooding*
Heat	High temperatures	*Heat exhaustion*
Heatwave	Extended hot weather	*Record-breaking heat*
Ice	Frozen precipitation	*Slippery surfaces*
Rain	Liquid precipitation	*Persistent showers*
Freezing	Below 0°C conditions	*Frost formation*
Snow	Frozen precipitation	*Accumulating snowfall*
Snowstorm	Intense snowfall	*Reduced visibility*
Storm	Severe weather event	*Strong winds/rain*
Thunder	Sound from lightning	*Loud rumbling*
Torrential	Extreme rainfall	*Flooding risk*

Table 7: Primary weather types and their definitions.

Ecological, Financial, Human Health, and Political — following previous studies (Imran et al., 2016a).

To ensure high-quality and consistent annotations, the task was conducted using a set of specific instructions reviewed by meteorological experts. The annotation guideline and the categories definition are provided in Table 14 and Table 15, respectively.

Notably, the same instruction guidance is contained within the prompts for LLMs in Appendix C to perform impact classification, following a binary output approach for each category.

B.2.1 Multi-Impact Labeling Format

Annotators are tasked with determining whether an article includes descriptions that correspond to the impact categories defined in Table 15. Each article is assigned a label based on the presence or absence of relevant descriptions.

- **1** – At least one mention of the relevant topic is identified.
- **0** – No relevant description is identified.

Special Case When an article describes multiple types of impact, each mentioned impact category is labelled as "1".

Dataset Statistics and Article Topics

To provide additional transparency regarding the dataset used in our analysis, we include detailed statistics of the included articles. The average number of tokens per article is 2987.4 in long-context settings and 781.3 in mixed-context settings.

Furthermore, we categorized the articles based on their associated weather-related topics. The table below shows the distribution of topics across the corpus:

A Special Case Example

Input Text: "Severe Storm Wreaks Havoc on Coastline. Bayport, September 15. A violent tempest swept the eastern seaboard last night, leaving a trail of devastation in its wake. The cargo vessel Harbor Star collided with a fishing trawler in the churning waters, capsizing the smaller craft and claiming three lives. Fierce winds reduced docks and piers to splinters, bringing commercial shipping to a standstill. The storm's toll is estimated to exceed $200,000, with Bayport Textile Mills filing for financial restructuring, placing 150 jobs in jeopardy. Hospitals are overwhelmed with storm-related injuries and illnesses, and emergency shelters are strained beyond capacity. The community now faces the arduous task of recovery."
Output: "Infrastructural: true, Agricultural: false, Ecological: false, Financial: true, Human Health: true, Political: false"

Table 8: A special case with multiple positive labels and is used for one-shot learning.

Weather	Count	Weather	Count
Snowstorm	34	Flood	22
Rain	32	Cold	21
Drought	31	Snow	17
Ice	30	Torrential	16
Storm	30	Blizzard	15
Thunder	29	Heatwave	13
Deluge	28	Heat	4
Freezing	28		

Table 9: Distribution of weather-related topics in the dataset. Each article may be assigned one or more topics based on content.

B.2.2 Instructions for Annotators

The instructions for annotators are provided in Table 14. The annotation process was conducted by members of a research group specializing in uncovering the history of a region's climate change through the regional historical weather records. Their expertise can ensure the accuracy and reliability of annotations.

C Instructions

C.1 Multi-Label Classification Instructions

The Multi-Label Classification instructions template in Table 16 is designed for both zero-shot and one-shot classification tasks.

- Zero-Shot: The model is given only the classification instructions and the input text.
- One-Shot for In-Context Learning: The model is provided with a demonstration for predicting a new sample. One example of demonstration we used is shown in Table 8.

C.2 Prompt Template for Multi-Label Text Classification

Table 16 presents the prompt designed to analyze historical newspaper texts and classify them into six distinct impact categories based on explicit mentions of weather-related events. The prompt is structured in alignment with the definitions provided in Table 15, which details the scope of each impact category, including Infrastructural, Agricultural, Ecological, Financial, Human Health, and Political impacts. The classification task is binary (true/false), requiring the model to identify whether the text explicitly mentions any of the defined impacts. The guidelines emphasize focusing on direct and immediate effects, ensuring that classifications are based solely on explicit references within the text. This prompt was used to evaluate multi-label classification models.

C.3 Prompt Template for Question Answering Ranking

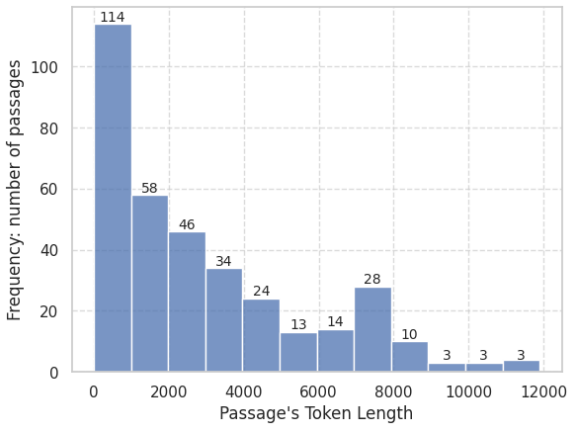
The ranking-based QA task consists of two key components: question generation (Mo et al., 2023) and candidate ranking (Meng et al., 2024). The prompts used for both stages are provided in Table 11 and Table 12, respectively.

D Dataset Statistic

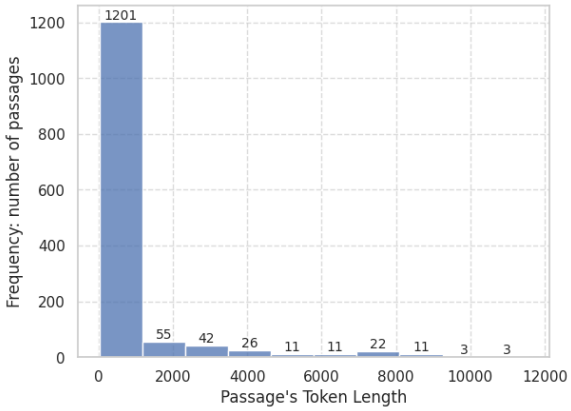
Figure 6 presents the token length distribution of passages in two versions of our dataset: (a) the Long Context dataset and (b) the Mixed Context dataset used for context-denoising evaluation.

The Long Context dataset (Figure 6a), which contains 350 articles, exhibits a broader distribution of passage lengths, with a significant portion exceeding 2000 tokens. While most passages are concentrated in the lower ranges, a noticeable number extend beyond 8000 tokens, demonstrating the dataset's emphasis on long-form content.

In contrast, the Mixed Context dataset (Figure 6b), which contains 1,386 articles, is heavily skewed toward shorter passages, with an overwhelming majority containing fewer than 2000 tokens. Only a small fraction of passages exceed 4000 tokens, highlighting the dataset's mixed nature, where shorter contexts are predominant.



(a) Long context dataset



(b) Mixed context dataset

Figure 6: The histogram shows the token length distribution (measured using the GPT-4 tokenizer) of articles in the two versions of our dataset.

E The Source of the Models

The models evaluated in this paper can be found as follows

1. GPT-4O, GPT-4 and GPT-3.5-TURBO are provided by OpenAI, the base model API document: https://platform.openai.com/docs/models
2. DEEPSEEK-V3-671B is upgraded the DEEPSEEK-CHAT, the base model API

Model	Infrastructural	Political	Financial	Ecological	Agricultural	Human Health	Average
Zero-Shot Performance							
GPT-4o	78.96 ↑ 0.18	84.44 ↓ 0.44	76.66 ↓ 0.95	78.39 ↑ 0.47	**84.44** ↑ 0.42	**80.40** ↓ 0.11	80.55 ↓ 0.07
GPT-4	72.33 ↑ 2.24	76.37 ↓ 2.37	59.65 ↑ 0.06	80.12 ↓ 3.26	79.25 ↑ 1.32	71.18 ↑ 0.82	73.15 ↑ 0.17
GPT-3.5-TURBO	77.52 ↑ 3.05	78.96 ↓ 0.67	69.45 ↓ 1.45	78.96 ↓ 1.82	80.69 ↑ 0.74	69.16 ↑ 1.70	75.79 ↑ 0.25
DEEPSEEK-V3-671B	**80.98** ↑ 0.45	**85.59** ↑ 1.84	**77.81** ↑ 0.48	81.56 ↓ 2.13	83.57 ↑ 1.00	77.23 ↑ 1.06	**81.12** ↑ 0.45
MISTRAL-24B-IT	76.30 ↓ 0.59	74.35 ↓ 3.78	69.45 ↓ 1.16	76.66 ↑ 0.20	82.42 ↑ 0.44	69.45 ↑ 0.26	74.77 ↓ 0.77
MIXTRAL-8X7B-IT	75.36 ↑ 2.16	80.29 ↓ 2.42	69.86 ↓ 4.06	**84.93** ↓ 3.32	78.26 ↑ 3.92	68.12 ↑ 1.99	76.14 ↑ 0.05
MISTRAL-7B-IT	77.23 ↑ 2.48	50.14 ↓ 1.00	34.58 ↓ 1.15	74.06 ↓ 3.49	72.05 ↑ 0.52	75.22 ↑ 0.49	63.88 ↑ 0.47
QWEN2.5-14B-IT	73.20 ↓ 2.91	70.03 ↓ 3.46	66.28 ↓ 2.28	66.86 ↓ 6.29	75.79 ↓ 7.22	70.32 ↓ 4.03	70.41 ↓ 4.37
QWEN2.5-7B-IT	72.05 ↑ 1.66	73.49 ↓ 8.35	67.72 ↓ 6.58	74.35 ↓ 10.06	73.49 ↓ 5.20	64.55 ↓ 2.84	70.94 ↓ 5.23
GEMMA-2-9B-IT	74.00 ↑ 3.49	61.14 ↓ 6.00	54.86 ↑ 2.26	66.29 ↓ 3.57	69.43 ↓ 5.43	64.00 ↓ 4.00	65.37 ↑ 0.47
LLAMA-3.1-8B-IT	73.49 ↑ 5.94	55.91 ↑ 0.66	62.25 ↓ 3.96	77.52 ↓ 2.66	77.81 ↑ 1.33	68.59 ↑ 4.55	69.26 ↑ 1.14
Average	75.58 ↑ 1.65	71.88 ↓ 2.36	64.42 ↓ 1.71	76.34 ↓ 3.27	77.93 ↓ 0.74	70.75 ↑ 0.00	72.82 ↓ 0.71
In-Context Learning (One-shot) Performance							
GPT-4o	79.25 ↓ 0.82	84.73 ↓ 1.02	74.64 ↑ 0.07	**79.83** ↓ 1.26	**85.01** ↓ 0.15	**80.12** ↑ 1.02	80.60 ↓ 0.36
GPT-4	70.89 ↑ 2.82	70.61 ↓ 0.90	61.10 ↓ 1.10	79.54 ↓ 1.25	79.25 ↑ 1.32	69.16 ↑ 2.27	71.76 ↑ 0.52
GPT-3.5-TURBO	73.49 ↑ 3.65	73.78 ↑ 4.51	55.33 ↑ 5.53	61.96 ↑ 9.47	77.23 ↑ 2.77	72.33 ↑ 0.24	69.02 ↑ 4.36
DEEPSEEK-V3-671B	**80.40** ↑ 1.03	**85.88** ↓ 0.71	**78.67** ↓ 0.67	79.83 ↓ 2.40	83.29 ↑ 1.28	77.81 ↓ 0.38	**80.98** ↓ 0.08
MISTRAL-24B-IT	76.88 ↓ 0.02	81.21 ↓ 0.64	76.01 ↓ 2.87	78.61 ↓ 1.18	80.06 ↑ 1.94	71.10 ↓ 1.39	77.31 ↓ 0.69
MIXTRAL-8X7B-IT	70.59 ↑ 4.12	77.49 ↓ 1.90	62.70 ↑ 3.77	71.38 ↑ 6.97	74.43 ↑ 1.75	65.81 ↑ 7.43	70.40 ↑ 3.52
MISTRAL-7B-IT	73.70 ↑ 1.73	69.08 ↑ 1.21	44.22 ↓ 1.36	71.68 ↓ 4.82	67.92 ↓ 8.78	70.52 ↓ 2.23	66.19 ↓ 2.38
QWEN2.5-14B-IT	76.08 ↓ 0.37	76.66 ↓ 4.37	71.76 ↓ 0.33	68.01 ↓ 3.72	78.96 ↑ 1.61	72.62 ↑ 1.09	74.02 ↓ 1.18
QWEN2.5-7B-IT	69.45 ↑ 0.84	82.42 ↓ 0.71	63.11 ↓ 5.68	62.82 ↓ 8.02	81.56 ↓ 3.27	60.52 ↑ 0.62	69.81 ↓ 2.48
GEMMA-2-9B-IT	71.14 ↑ 0.62	67.14 ↓ 3.09	50.29 ↓ 2.31	60.57 ↓ 2.15	74.86 ↑ 2.89	68.00 ↓ 0.79	65.33 ↑ 0.62
LLAMA-3.1-8B-IT	74.06 ↑ 3.08	64.55 ↑ 2.02	55.91 ↑ 0.38	76.95 ↓ 6.09	74.64 ↓ 0.93	69.16 ↑ 0.55	69.05 ↑ 0.19
Average	74.17 ↑ 1.52	75.78 ↓ 0.51	63.07 ↓ 0.42	71.93 ↓ 1.31	77.93 ↑ 0.04	70.65 ↑ 0.77	72.26 ↑ 0.19

Table 10: Accuracy by percentage results of zero-shot and one-shot evaluation categorized on six impacts and two context length settings. The used notations are the same as Table 1.

Question Generation Template

```
Given the following passage about
{row['Weather']}, generate a
single, focused question that
meets these criteria:
1. Can be answered using ONLY the
information in this passage
2. Focuses on the {impact_str}
impacts mentioned
3. Is detailed and specific to
this exact situation
4. Requires understanding the
passage's unique context
5. Cannot be answered by other
similar passages about
{row['Weather']}
Passage: {row['Text']}
```

Table 11: Instruction used to generate Questions in the ranking-based QA task.

documents: https://api-docs.deepseek.com/

3. MIXTRAL-8X7B-IT[2], MISTRAL-24B-IT[3], MISTRAL-7B-IT[4], LLAMA-3.1-8B-IT[5], QWEN2.5-14B-IT[6], QWEN2.5-7B-IT[7] and GEMMA-2-9B-IT[8], are base models weights from the Huggingface website: https://huggingface.co/

F Computation Cost

For the large proprietary models (e.g., GPT-4o), conducting a one-time evaluation on our WXImpactBench costs approximately $3 for multi-label classification tasks and $5.5 for ranking-based QA tasks. For all open-source models, evaluations were performed on a system with two NVIDIA A6000 (32GB) GPUs. The relatively modest computational requirements highlight the accessibility of our benchmark for researchers with limited computational resources, while still enabling comprehensive evaluation of state-of-the-art models

[2] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[3] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[7] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[8] https://huggingface.co/google/gemma-2-9b-it

QA Ranking Task Prompt

```
{
"role": "system",
"content": "You are an expert that
ranks passages based on their
relevance to a given query.
  The most relevant passage should
answer the query"
},{
"role": "user",
"content": f"Query: {query} Rank
the following passages
[{start_idx+1} to
{start_idx+len(passages)}] by
relevance."
}
for i, passage in
enumerate(passages):
messages.extend([
"role": "user", "content":
f"[start_idx+i+1] passage",
"role": "assistant", "content":
f"Received passage
[start_idx+i+1]"
])
{
"role": "user",
"content": "Provide ranking as
numbers separated by '>',
  e.g., [3] > [1] > [2] > [5] >
[4]. No explanation needed."
}
```

Table 12: Instruction used in Ranking-based QA task.

G Additional Experimental Results

LLMs might be more effective in historical narratives. Table 13 presents the performance of the evaluated LLMs on WXIMPACTBENCH across historical and modern impact categories in the one-shot setting. The results are categorized based on six societal impact dimensions with varying context lengths.

H Annotation Guidelines

To ensure a high-quality evaluation of historical weather impact analysis, we developed a structured annotation framework for meteorology experts. The goal of this annotation is to create a reliable benchmark for assessing the ability of LLMs to understand and classify disruptive weather-related societal and environmental impacts. The detailed annotation guidelines are provided in Table 14, outlining the task objectives, category definitions, and better practices for identifying and classifying weather impacts in historical texts.

Model	Historical	Modern
Cases	200 (504)	150 (882)
GPT-4O	**70.24** ↑ 1.09	**69.05** ↑ 0.97
GPT-4	62.77 ↑ 1.81	50.89 ↑ 5.85
GPT-3.5-TURBO	60.26 ↑ 1.18	52.63 ↑ 3.91
DEEPSEEK-V3-671B	65.77 ↑ 3.68	67.83 ↑ 0.68
MISTRAL-24B-IT	63.51 ↑ 4.16	61.27 ↓ 0.64
MIXTRAL-8X7B-IT	49.17 ↑ 8.83	41.99 ↑ 9.01
MISTRAL-7B-IT	53.11 ↑ 3.75	48.16 ↓ 0.92
QWEN-2.5-14B-IT	60.66 ↑ 2.42	59.83 ↑ 0.84
QWEN2.5-7B-IT	52.74 ↑ 2.60	48.84 ↓ 1.76
GEMMA-2-9B-IT	57.82 ↑ 0.86	51.82 ↑ 2.73
LLAMA-3.1-8B-IT	55.06 ↑ 4.89	50.75 ↑ 4.03
Average	59.19 ↑ 3.2	54.82 ↑ 2.25

Table 13: F1-score performance across historical and modern impact categories in the one-shot setting. The used notations are the same as Table 1.

Instructions For Annotators

Annotation Guidelines for Historical Weather Impact Analysis

This document provides comprehensive guidelines for annotators who analyze historical newspaper articles describing disruptive weather events. The primary objective is to identify and categorize six distinct impact categories within each text. This analysis will facilitate comparisons with the performance of large language models in understanding weather-related impacts across various societal and environmental dimensions.

Task Overview

Annotators will examine historical newspaper articles documenting disruptive weather events. The analysis requires the identification of impacts across six categories: infrastructural, agricultural, ecological, financial, human health, and political. Please refer to Table 15 for the definitions of these categories.

Note: While specific examples are provided for each impact category, annotators should apply their meteorological expertise to identify and classify impacts beyond these examples, maintaining a comprehensive analytical approach.

If you have any questions, please feel free to contact us. Thank you for your invaluable support!

Table 14: Instructions for annotators

Category	Definition
Infrastructural Impact	Examines weather-related damage or disruption to physical infrastructure and essential services. Includes structural damage to buildings, roads, and bridges; disruptions to transportation (e.g., railway cancellations, road closures); interruptions to utilities (e.g., power, water supply); failures in communication networks; and industrial facility damage. Both immediate physical damage and service disruptions should be considered.
Agricultural Impact	Focuses on weather-related effects on farming and livestock management. Includes crop yield variations; direct damage to crops, timber, or livestock; modifications to farming schedules; disruptions to food production and supply chains; impacts on farming equipment; and changes in agricultural inputs (e.g., soil conditions, water availability, fertilizers, animal feed). Both immediate and long-term effects should be considered.
Ecological Impact	Examines effects on natural environments and ecosystems. Includes changes in biodiversity; impacts on wildlife populations and behavior; effects on non-agricultural plant life; habitat modifications (e.g., forests, wetlands, water bodies); changes in hydrological systems (e.g., river levels, lake conditions); and urban plant life impact. Immediate environmental changes should be prioritized.
Financial Impact	Analyzes economic consequences of weather events. Includes direct monetary losses; business disruptions requiring financial intervention; market fluctuations; impacts on tourism and local economies; and insurance claims or economic relief measures. The focus should be on explicit financial impacts rather than inferred consequences.
Human Health Impact	Examines both physical and mental health effects. Includes direct injuries or fatalities (including cases where one or more casualties are explicitly mentioned); increased risks of weather-related illnesses; mental health consequences (e.g., stress, anxiety); impacts on healthcare accessibility; and long-term health implications. Both short-term and chronic health effects should be considered.
Political Impact	Evaluates governmental and policy responses to weather events. Includes government decision-making and policy changes; shifts in public opinion or political discourse; effects on electoral processes; international aid and relations; and debates on disaster preparedness and response. Both direct political reactions and policy implications should be analyzed.

Table 15: Impact categories and their definitions

Multi-Label Classification Task: Zero-Shot Instruction Template

Given the following historical newspaper text:
"{instruction}"

Analyze the text and provide a binary classification (respond ONLY with 'true' or 'false') for each impact category based on explicit mentions in the text. Follow these specific guidelines

1. **Infrastructural Impact**: Classify as 'true' if the text mentions any damage or disruption to physical infrastructure and essential services. This includes structural damage to buildings, roads, or bridges; any disruptions to transportation systems such as railway cancellations or road closures; interruptions to public utilities including power and water supply; any failures in communication networks; or damage to industrial facilities. Consider only explicit mentions of physical damage or service disruptions in your classification.

2. **Agricultural Impact**: Classify as 'true' if the text mentions any weather-related effects on farming and livestock management operations. This includes yield variations in crops and animal products; direct damage to crops, timber resources, or livestock; modifications to agricultural practices or schedules; disruptions to food production or supply chains; impacts on farming equipment and resources; or effects on agricultural inputs including soil conditions, water availability for farming, and essential materials such as seedlings, fertilizers, or animal feed.

3. **Ecological Impact**: Classify as 'true' if the text mentions any effects on natural environments and ecosystems. This includes alterations to local environments and biodiversity; impacts on wildlife populations and behavior patterns; effects on non-agricultural plant life and vegetation; modifications to natural habitats including water bodies, forests, and wetlands; changes in hydrological systems such as river levels and lake conditions; or impacts on urban plant life.

4. **Financial Impact**: Classify as 'true' if the text explicitly mentions economic consequences of weather events. This includes direct monetary losses; business disruptions or closures requiring financial intervention; market price fluctuations or demand changes for specific goods; impacts on tourism and local economic activities; or insurance claims or economic relief measures. Focus only on explicit mentions of financial losses or fluctuations.

5. **Human Health Impact**: Classify as 'true' if the text mentions physical or mental health effects of weather events on populations. This includes direct injuries or fatalities (including cases where zero or more casualties are explicitly mentioned); elevated risks of weather-related or secondary illnesses; mental health consequences such as stress or anxiety; impacts on healthcare service accessibility; or long-term health implications.

6. **Political Impact**: Classify as 'true' if the text mentions governmental and policy responses to weather events. This includes government decision-making and policy modifications in response to events; changes in public opinion or political discourse; effects on electoral processes or outcomes; international relations and aid responses; or debates surrounding disaster preparedness and response capabilities.

Note:
- Return 'false' for any impact category that is either not present in the text or not related to weather events
- Base classifications on explicit mentions in the text
- Focus on direct impacts rather than implications
- Consider immediate and direct effects

Answer only once in the following format:
Infrastructural: true/false
Agricultural: true/false
Ecological: true/false
Financial: true/false
Human Health: true/false
Political: true/false

Table 16: Multi-Label Classification instructions template used as the Zero-Shot prompt.