Beyond Consensus: Diagnosing Coverage, Majority Bias, and Grounding Equity in High-Stakes Multi-Document Diversity Summarization

Jane Arleth dela Cruz **Iris Hendrickx** **Martha Larson**
Center for Language and Speech Technology
Center for Language Studies
Radboud University, Nijmegen, Netherlands
{janearleth.delacruz, iris.hendrickx, martha.larson}@ru.nl

Abstract

Multi-document diversity summarization (MDDS) in high-stakes domains such as crisis response demands fundamentally different capabilities than consensus-based summarization. While standard summarization focuses on fluency and faithfulness, high-stakes tasks would also require: maximizing coverage of divergent information, identifying unique information clusters, and providing verifiable citations to source documents. Existing evaluation frameworks fail to capture these task-specific requirements. Our study introduces a comprehensive evaluation framework for high-stakes MDDS, validated on a multi-source crisis dataset that we developed covering ten crisis events. We propose novel metrics, Comprehensive Coverage Score, Frequency-Salience Correlation, and Citation-Frequency Correlation, to diagnose critical failure modes in six off-the-shelf large language models (LLMs). Our results allowed us to identify three critical failure modes of LLMs: (1) a consensus trap in coverage: where models capture high-level information clusters, but miss specific, actionable details needed for decision-making; (2) majority bias, where more frequent information is prioritized over rare but important ones and; (3) a verifiability gap, where citations of rare claims are less reliable.

1 Introduction

Large language models (LLM) have demonstrated remarkable capabilities to generate fluent and coherent text. In some settings, they can rival well-paid freelancers in writing news summaries (Zhang et al., 2024a). However, the rapid adoption of LLMs for automatic summarization raises concern on whether current evaluation practices adequately capture the full-scope of requirements of different summarization tasks. LLMs are becoming the default tool for automatic multi-document summarization, even being introduced and deployed in high-stakes domains (Lei et al., 2025; Heddaya et al., 2025; Tang et al., 2023). In these settings, it is essential to evaluate not only fluency and faithfulness, but also task-specific requirements.

Huang et al. (2024) differentiate summarization approaches that require consensus of information and those that require the breadth of differing viewpoints. For example, movie review summaries, news summaries, and executive briefings summaries would require the most frequently occurring elements; the goal is to find where the sources most agree upon (Fabbri et al., 2019). However, in other contexts, such as crisis response and investigative journalism, the requirement shifts towards coverage: capturing all the diverse information from sources or different perspectives (Huang et al., 2024; McCreadie and Buntain, 2023; Tang et al., 2025). Huang et al. (2024) formalized **multi-document diversity summarization** (MDDS) as the task of generating a summary that faithfully captures all the diverse information presented in the input sources.

Current evaluation of MDDS remains insufficient for high-stakes domains. For example, Huang et al.'s 2024 evaluation relies primarily on LLM-as-a-judge, which reduces interpretability and introduces potential bias, as previous studies of LLM-evaluating-LLM have shown (Ye et al., 2025; Gu et al., 2026). Furthermore, existing literature has largely focused on measuring faithfulness (whether summaries contain hallucinations) and general fluency (whether summaries are coherent). While these dimensions are still necessary, they are insufficient for high-stakes MDDS where the consequences of incomplete coverage are real (resource allocation and delayed response), i.e., a potential cause of harm.

In addition, existing datasets and annotation frameworks are inadequate to evaluate the high-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4831–4845
November 6-10, 2026 ©2026 Association for Computational Linguistics

stakes MDDS requirements in real-world settings. For example, Laban et al. (2024) used an LLM to generate a dataset covering 92 summary haystacks of insights, however, their dataset is both controlled for insight count (instead of natural frequency distribution) and lacks representative information per insight group. These design choices limit the ability to diagnose the full range of requirements in the high-stakes task.

The primary challenge is that LLMs are not neutral information processors (Gallegos et al., 2024; Steen and Markert, 2024). In a crisis scenario, LLMs may fall into a **consensus trap**—capturing high-level, frequently mentioned themes while systematically omitting actionable details from rare sources (e.g., where people are needing rescue or which shelters are full). Current summarization metrics fail to capture **majority bias**—the tendency to prioritize more frequent information. Hence, we focus on how LLMs maximize coverage across the frequency spectrum of information sources (e.g., official evacuation call from ten (10) news sources vs. two (2) posts of people trapped) and how they attribute information to input sources.

We argue that high-stakes MDDS requires a fundamentally different set of standards, and we investigate this in the crisis summarization task. Note that we focus on query-based (focusing on human impact) summarization. We identify three core requirements for this task: (1) maximizing information coverage by integrating all divergent, and possibly conflicting information that characterizes the sources, (2) grouping redundant information into unique information clusters and identifying a representative information item per cluster, and (3) attributing the summary by citing all the source documents of all summary information items. We refer the requirement (3) as **grounding equity**. Grounding equity is when attribution is applied consistently, regardless of frequency, which means that rare claims are as reliable (grounded) as frequent claims.

In this work, we tackle the limitations and address these requirements by developing an evaluation framework for the high-stakes MDDS task. Our main contributions are:

- **Evaluation framework for high-stakes MDDS for crisis events:** we develop an annotation framework that allows for the task-specific requirements of crisis summarization. We applied the framework to develop a dataset aligned for the evaluation. To do this, we leverage the existing CrisisFacts dataset (McCreadie and Buntain, 2023) to construct a task focused on *human impact and welfare* information as the **target concept**, across ten crisis events using data from multiple platforms i.e., news, X (formerly Twitter), Facebook, and Reddit.
- **Task-specific novel evaluation metrics for MDDS:** we introduce metrics that measure the specific requirements for MDDS in high-stakes settings: Comprehensive Coverage Score, Frequency-Salience Correlation and Citation-Frequency Correlation . These metrics measure diverse information coverage, majority bias detection, and verifiability going beyond LLM-as-a-judge evaluations used in previous MDDS work.
- **Comprehensive Analysis of LLMs in crisis MDDS task:** we investigate the performance of six closed and open-sourced LLMs.

Our results allowed us to identify three critical failure modes of LLMs in the high-stakes MDDS task. First, we observed a consensus trap in coverage: while models capture high-level information clusters, they struggle to report the specific, actionable details needed for decision-making. Second, LLMs exhibit majority bias, prioritizing information frequency in summary creation. Third, we identified a verifiability gap, where citation reliability is also influenced by frequency.

2 Related Work

We position our work within research areas on evaluation metrics for summarization, MDDS, dataset limitations, and bias detection.

Evaluation Metrics for Multi-Document Summarization. Traditional evaluation relies heavily on n-gram overlap metrics like ROUGE, which measures the lexical overlap between system and reference summaries (Lin, 2004). Recent work has addressed n-gram-based limitations by introducing contextual embedding-based metrics, such as BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), but they still struggle particularly in multi-document or longer context settings (Fabbri et al., 2021; Guo and Vosoughi, 2023; Wang et al., 2023). Faithfulness metrics assess whether summaries introduce

hallucinated content not present in the source (Maynez et al., 2020; Laban et al., 2022; Tang et al., 2024; Zha et al., 2023), but they evaluate what is present in the summary, not what is missing, making them blind to incomplete coverage of low-frequency claims.

Diverse Information Coverage and MDDS. Recent work has formalized MDDS to address coverage needs (Huang et al., 2024; Zhang et al., 2024b). Both Huang et al. (2024) and Laban et al. (2024) proposed LLM-as-a-judge metrics for coverage, but neither distinguishes between identifying broad clusters and reporting specific actionable details. Our work extends this by proposing Comprehensive Coverage Score designed to take into account both breadth and depth of coverage.

A parallel coverage challenge is maintaining focus on the target information need. We introduce Coverage Disparity (non-target concept score) to quantify how much a summary covers tangential topics (e.g., animal welfare or dogs in shelters). While faithfulness metrics (Laban et al., 2022; Tang et al., 2024; Zha et al., 2023) ensure that summaries do not hallucinate information, they fail to catch source-supported non-target concepts. In high-stakes settings, a summary that spends 15% of its output on animal welfare instead of human impact diminishes real-world utility.

Limitations of Existing Datasets for MDDS. Existing MDDS and crisis summarization datasets are inadequate for diagnosing frequency-aware coverage, majority bias, and grounding equity. The original CrisisFacts Dataset (McCreadie and Buntain, 2023) provides real-world multi-platform crisis streams across crisis events but lacks manually annotated reference summaries. They developed gold summaries based on pooled task submissions (consensus-based), where NIST assessors provided relevance and redundancy judgments (McCreadie and Buntain, 2023). This lacks claim/information-level frequency data, unique cluster annotations, and citations. DiverseSumm (Huang et al., 2024) is news only and provides answer clusters without annotating answer cluster representatives. Laban et al. (2024) provided a framework to synthesize summary haystacks containing pre-selected insights. However, it lacks the natural frequency distribution present in real crisis data, thus, cannot diagnose majority bias.

Detecting Bias in Summarization. LLM summarization inherits social and content-selection biases (Gallegos et al., 2024; Zhou et al., 2025; Steen and Markert, 2024; Zhang et al., 2024b). Studies have focused on identifying and measuring social biases in language models, particularly related to gender and race (Steen and Markert, 2024; Anantaprayoon et al., 2024; Liang et al., 2023), as well as measuring position bias in summarization (Chhabra et al., 2024; Olabisi and Agrawal, 2024; Ravaut et al., 2024). However, majority bias, remains underexplored. Huang et al. (2024) noted that long-context LLMs had better coverage for more frequently occurring answers; however, they did not operationalize this observation into a metric.

Attribution and Verifiability. As LLMs are increasingly introduced and deployed in high-stakes domains, verifiable attribution is critical (dela Cruz et al., 2025). Benchmarks were introduced to evaluate LLMs' ability to ground their output with citations (Li et al., 2024; Clark et al., 2023). Li et al. (2024) and Laban et al. (2024) assessed overall citation quality with precision, recall, and F1-score metrics, but they treat all facts or clusters uniformly and did not examine citation reliability on a granular level. In high-stakes MDDS, end users would most need reliable citations for rare, therefore, harder to verify claims (McCreadie and Buntain, 2023). In our work, we introduce Citation-Frequency correlation, a novel metric that measures grounding equity.

3 Task Formulation

We study high-stakes MDDS in a crisis setting, where each input document is a time-stamped post from one of four different online platforms, i.e., Facebook, X, Reddit, and News (McCreadie and Buntain, 2023).

New CrisisFacts Summarization Task. We construct a new high-stakes MDDS task grounded in the CrisisFacts dataset (McCreadie and Buntain, 2023) to reflect the decision-making needs of disaster management. The CrisisFacts task broadly asked, *"What critical new developments have occurred today that I need to know about?"*, which was accompanied by up to 52 queries covering everything from infrastructure damage to emergency response. (McCreadie and Buntain, 2023).

To isolate the high-stakes requirements of MDDS, we significantly narrow this scope. For each event-day, we focus exclusively on direct reports of *human impact and welfare*. This reduced the scope to the data associated with seven

targeted queries: *How many people are affected (evacuated)?, How many people are in shelters?, etc.* The full list is found in Appendix A.1.

Given an input batch of event-day posts, an ideal LLM summary should satisfy the core requirements of high-stakes MDDS:

1. **Information Coverage:** Identify all posts relevant to *human impact and welfare*.
2. **Information Clustering:** Group relevant posts into unique information clusters that report similar information. Identify a representative post for each cluster.
3. **Grounding Equity:** Cite all the supporting source posts for each information cluster, regardless of frequency of information.

No existing annotation framework currently supports the evaluation of the three core requirements simultaneously in real-world crisis-settings. For example, the reference summaries for CrisisFacts task was based on pooled system submissions where assessors provided relevance and redundancy judgments (McCreadie and Buntain, 2023). Their framework does not group redundant information clusters, select representative spans or enforce source citation. Therefore, a new annotation framework is necessary for evaluation.

Manual Gold Summary Construction. We constructed a new gold standard summary based on the original CrisisFacts task dataset (McCreadie and Buntain, 2023). We performed manual annotation on batches of 100 posts per event-day, grouped by timestamp. The manual annotation process consisted of three steps: (1) Relevance annotation, (2) Redundancy check and cluster creation and (3) Representative post selection. A representative post is selected as the single most informative post in each cluster (a cluster contains highly similar information but differ in specificity, phrasing, or completeness). This annotation process reflects how emergency responders and disaster managers receive updates in streaming batches of online media data, not as a single document. The 100-post batch size was chosen to be sufficiently large to reflect realistic manual annotation difficulty—where LLM assistance can be helpful and where it is within context-range so LLMs can still be reliable. The full annotation process description is in Appendix A.2.

The final **manual gold summary** consists of the resulting list of representative posts (gold) which are extractive, non-repeating, highly relevant posts with the citation of all sources.

4 Experimental Setup

4.1 Dataset

As previously mentioned, we created our dataset by leveraging the **CrisisFacts Dataset** (McCreadie and Buntain, 2023). We run our experiments on ten (10) crisis events from the CrisisFacts dataset that had open access. We employed sampling and preprocessing steps to arrive at the final dataset as described in Appendix A.1. Our final dataset includes 11,770 documents from 63 days of crisis (Table 4). We release the dataset in and annotation protocol in Dela Cruz et al. (2026).

4.2 Models

We chose commonly used off-the-shelf LLMs. We used GPT 4o-mini (OpenAI, 2024), GPT 5-mini (OpenAI, 2025), Llama 3.1-8B Instruct (Llama Team, 2024), Llama 3.2-3B Instruct (Llama Team, 2024), and Mistral NeMo (Mistral AI Team, 2024), and Qwen 3-4B (Yang et al., 2025). They were chosen because they are commonly used by both researchers and practitioners. Details about the models are in Appendix B.

4.3 Evaluation Metrics

Table 1 shows the overview of the evaluation metrics and the key questions they address. Our evaluation framework directly addresses the three core requirements for high-stakes MDDS. Figure 1 shows a walkthrough example LLM summary evaluation for an event-day batch.

Requirement 1: Information Coverage

Factual Accuracy (Table 1, row 2). We evaluate the factual accuracy of the LLM summaries using the fact matching metrics from CrisisFacts (McCreadie and Buntain, 2023). We measure Comprehensiveness (recall-oriented overlap between LLM summary and the manually annotated human impact relevant set) and Redundancy (precision-oriented ratio of matched LLM summary and all generated summary), and report the F1-score. Our evaluation focuses on matching the presence of information via the LLM's generated citations and does not evaluate full faithfulness (text-matching).

Comprehensive Coverage Score (Table 1, row 2). We measure coverage by how well the summary includes all relevant information. Cov-

Requirement	Evaluation Aspect (Metric)	Key Question Answered
Maximizing Information Coverage	**Factual Accuracy** (Factual F1-score)	To what extent does the summary capture verified human impact information while minimizing redundancy?
	Coverage (Comprehensive Coverage Score)	Does the summary balance *breadth* (finding all relevant information clusters) with *depth* (reporting the representative post)?
	Coverage Disparity (non-target concept score)	How much summary content is diverted to tangential topics (e.g., animal welfare) instead of the core information need?
Information Clustering	**Majority Bias** Frequency-Salience Correlation (ρ)	Are high-frequency information clusters systematically prioritized while marginalizing rare but critical ones within the relevant set?
Attributing to Sources	**Verifiability** Citation Score (F1-score)	Are summary clusters correctly attributed to their source documents?
	Grounding Equity Citation-Frequency Correlation (ρ)	Are citations for rare information clusters less reliable (more prone to hallucination) than for frequent ones?

Table 1: Summary of evaluation metrics. We assess performance across three core requirements of high-stakes multi-document diversity summarization (MDDS): (1) maximizing information coverage, (2) information clustering and selecting representative information, and (3) attributing to sources.

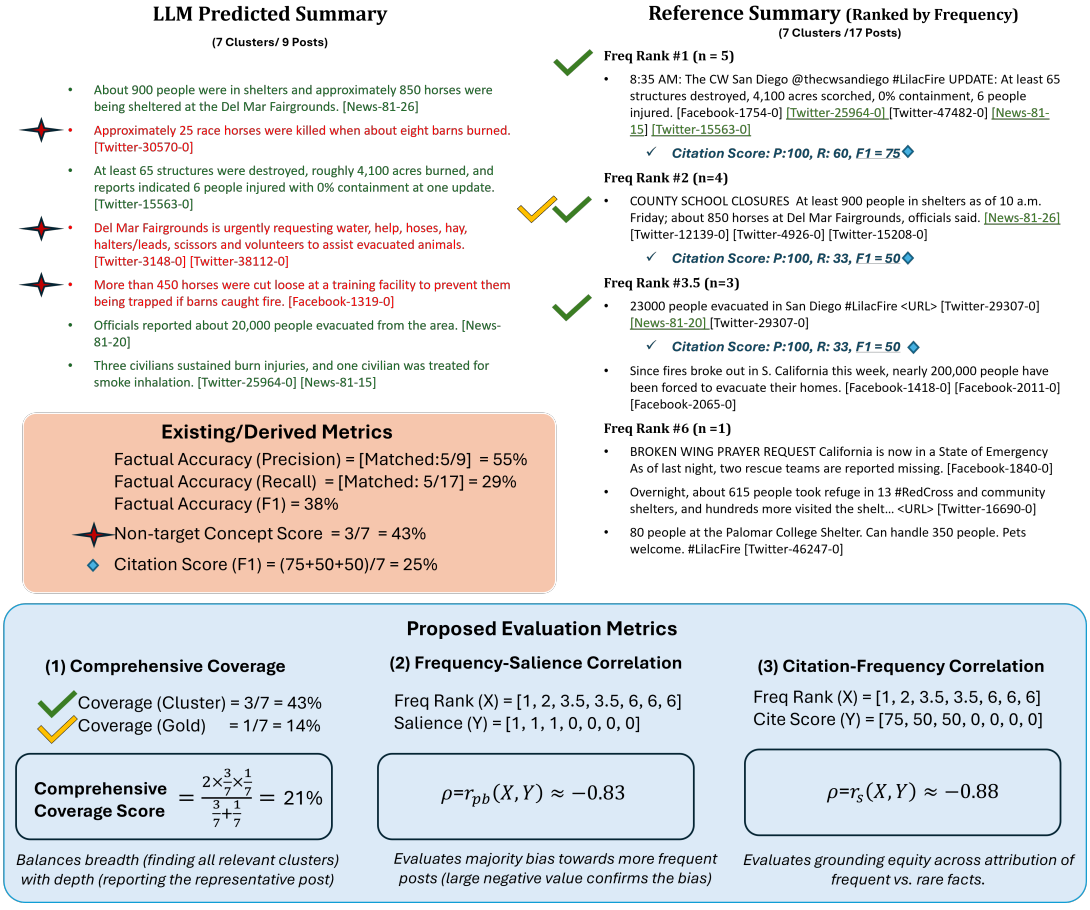


Figure 1: Walkthrough of the evaluation pipeline comparing an LLM predicted summary (left, GPT-5-mini) against reference human impact summary (right) for an event-day batch of Lilac Wildfire event on February 8, 2017. Green check marks (✔) denote covered clusters (Coverage (Cluster)), the gold check mark (✔) marks representative gold post match (Coverage (Gold)), stars (★) indicate non-target tangential posts (about animal welfare), and blue diamonds (◆) indicate Citation F1 Scores. The blue box show the derivation of our proposed metrics: (1) Comprehensive Coverage Score measures the balance between breadth and depth of coverage; (2) Frequency-Salience Correlation ($\rho = -0.83$) measures majority bias toward frequent clusters; and (3) Citation-Frequency Correlation ($\rho = -0.88$) diagnoses the verifiability gap on rare facts.

erage (Cluster) scores 1 per covered cluster (at least one item covered), 0 otherwise. Coverage (Gold) scores 1 per fully covered representative post (Gold) per cluster. We introduce a novel metric, Comprehensive Coverage Score, which is the harmonic mean of both coverage scores (0-1 range). It balances the breadth vs. depth of the information covered.

Coverage Disparity (Table 1, row 3). We introduce non-target concept score, to quantify the degree to which the summary is distracted by non-human impact information that are tangential to core information need. We classify each summary information item as "Target" or "Non-Target", using an easy-to-implement keyword-based classification approach based on an ontology of target and non-target concepts. While highly efficient, this method may be less robust to synonyms and paraphrasing.

Requirement 2: Information Clustering

Frequency-Salience Correlation (Table 1, row 4). We measure the relationship between frequency of posts in a cluster (cluster size) and selection salience within the human impact relevant set. Our scope reflects the dataset's annotation process, which only clustered redundancy within the relevant set, while the much larger non-relevant set was not clustered.

We introduce the Frequency-Salience correlation (ρ) metric. For each event-day batch (100 posts), let $\{f_i\}$ be the set of human-impact clusters in that batch. For each cluster f_i:

- Frequency Rank(f_i): Rank of the cluster (f_i) by redundancy count (1 = most frequent).
- Salience(f_i): 1 if the summary contains ≥ 1 item from f_i, 0 otherwise; this was also referred to as Coverage (Cluster).

For each event-day batch, we compute:

$$\rho = \text{Point-Biserial}(\{\text{Salience}(f_i)\}, \{\text{Rank}(f_i)\})$$

Values where $\rho < 0$, indicates *majority bias* (higher-ranked clusters more likely selected). A high negative correlation ($\rho < 0$) indicates majority bias, where more frequent human-impact posts in the source are systematically more likely to be selected. A low correlation suggests that selection relies more by other factors such as relevance or diversity.

Requirement 3: Attributing to Sources

Citation Score (Table 1, row 5). To ensure verifiable summaries, we use Citation Score, adapted from Laban et al. (2024), to measure how effectively the LLM grounds its summary statements in source documents. We compute the F1-score of precision and recall between the generated citations for each cluster and the reference cluster citations. The final Citation Score is reported as the macro-average F1-score across all reference clusters.

Citation-Frequency Correlation (Table 1, row 6). To assess grounding equity, we compute the Citation-Frequency correlation: the Spearman correlation (ρ) between a cluster's frequency rank and its Citation Score. Negative mean ρ indicates that citations for rare posts are less reliable than for frequent ones.

4.4 Prompts

Our prompts were constructed with reference to our annotation framework and the original CrisisFacts task. We based our prompt format on Laban et al. (2024), as our desired output was similar. Output summaries are in bullet-points with citation to source ID's in bracket format, e.g., [News-3-4][Twitter-2-7]. Lastly, we restricted the output to be a maximum of seven (7) bullets which was the maximum number of clusters in the reference summaries, and as length is a known confounding factor in summarization evaluation (Guo and Vosoughi, 2023; Liu et al., 2023; Laban et al., 2024). The full prompt is found in Appendix C.

5 Results

The summary of the performance of the LLMs in high-stakes MDDS is found in Table 2.

Factual Accuracy

The results of the factual accuracy evaluation of the LLMs in MDDS are found in Table 3. We observed that the crisis summarization task is difficult for LLMs with comprehensiveness scores ranging from 34.15 to 46.37, redundancy scores ranging from 27.47 to 42.17 and F1-score values of 29.21 to 37.15.

Comprehensive Coverage

Figure 2 summarizes the LLMs performance in MDDS with respect to the coverage metrics. We compute the coverage per batch of 100-post event-days and average these across all the summaries.

Model	Diverse Information Coverage			Majority Bias	Verifiability	
	Factual Accuracy (F1-score) ↑	Comprehensive Coverage Score ↑	Coverage Disparity (non-target) ↓	Frequency-Salience Score (ρ)	Citation Score (F1-score) ↑	Citation-Frequency Score (ρ)
Llama 3.2-3B	30.11	37.17	3.13	-0.42	35.11	-0.16
Llama 3.1-8B	36.09	44.20	3.85	-0.48	43.40	-0.19
Qwen 3 4B	29.21	39.08	5.55	-0.46	36.47	-0.19
Mistral NeMo	37.10	45.79	3.91	-0.52	43.29	-0.27
GPT 4o-mini	34.88	45.39	4.20	-0.47	43.34	-0.24
GPT 5-mini	37.15	54.58	4.36	-0.30	49.60	0.04

Table 2: Results of MDDS are reported based on the summary requirements: (1) Diverse Information Coverage is measured by Factual Accuracy (F1-score of matching all relevant facts), Comprehensive Coverage Score (balance of breadth and depth of information selection), Coverage Disparity (percentage of tangential information), (2) Majority Bias is measured by the Frequency-Salience Correlation (relationship between clusters' frequency rank and salience of information) and (3) Verifiability is measured by Citation Score (reliability of LLM citations) and Citation-Frequency Correlation (relationship between clusters' frequency rank and Citation Score).

Model	Comprehen-siveness (Recall)	Redundancy (Preci-sion)	F1-score
Llama 3.2-3B	34.15	33.78	30.11
Llama 3.1-8B	41.71	40.59	36.09
Qwen 3 4B	36.49	27.47	29.21
Mistral NeMo	42.73	42.17	37.10
GPT 4o-mini	42.31	36.56	34.88
GPT 5-mini	46.37	38.54	37.15

Table 3: Factual Accuracy: Performance of LLMs in MDDS, comparing LLM summary to the Human Impact (Relevant Set).

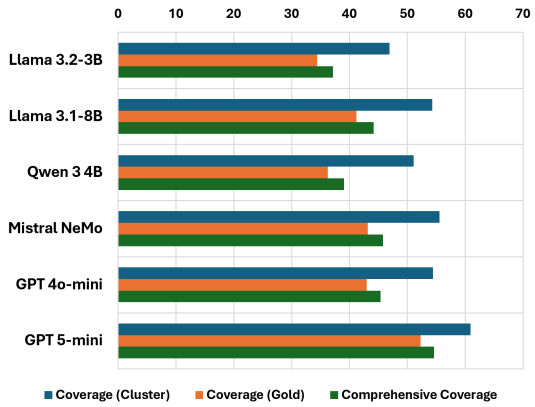


Figure 2: Coverage Scores: Performance of LLMs in MDDS, comparing LLM summary to the reference summary. Comprehensive Coverage is the harmonic mean of Coverage (Cluster): coverage to any item in a cluster and Coverage (Gold): coverage of the representative post in a cluster.

We observed that LLMs have higher Coverage (Cluster) scores than Coverage (Gold) scores, indicating finding at least one relevant post per cluster is relatively easier than identifying the representative post. The Comprehensive Coverage scores range from 37.17 to 54.58, with GPT 5-mini having the highest coverage (Table 2).

Coverage Disparity

We observe quite low average non-target concept scores that are all below 6% (Table 2), which means LLMs are not easily distracted with non-information needs. We further looked into each crisis event to see whether this is true for every event (see Figure 3). We observe that the non-target concept scores are highly dependent on the type of crisis event being summarized. For example, all the LLMs have struggles on Lilac Wildfire with non-target concept scores from 10.32% to 18.19% and Saddleridge Wildfire with non-target concept scores of 4.2% to 9.52%.

Majority Bias

We present the mean event-day batch frequency-salience correlation values of the LLMs in Table 2. There is a consistent negative correlation between the frequency rank of an information cluster and the cluster's selection in the LLM summary. This means that LLMs tend to prioritize selecting more frequent facts and exhibit majority bias.

Verifiability

Citation Scores (Table 2) account for both the precision and recall of the LLMs' attribution to input documents. GPT 5-mini has the highest Citation Score at 49.60, followed by GPT 4o-mini at 43.34.

As for grounding equity, we observed that there is a low negative mean event-day batch correlation between the Citation Score and the frequency rank of an information cluster for all the LLMs except for GPT 5-mini. These values reflect directional trends aggregated across event-day batches.

Event	Model					
	Llama 3.2	Llama 3.1	Qwen 3	Mistral Nemo	GPT 4o-mini	GPT 5-mini
001 Lilac Wildfire	10.32	12.96	14.81	12.17	13.23	18.19
002 Cranston Wildfire	2.08	1.98	7.14	3.17	2.78	3.77
003 Holy Wildfire	4.76	7.14	8.33	7.34	5.95	0
004 Hurricane Florence	0.51	1.19	2.04	0.68	1.19	2.38
005 Maryland Floods	0	0	0	0	0	0
006 Saddleridge Wildfire	6.55	4.2	8.33	8.33	9.52	4.76
007 Hurricane Laura	3.57	9	3.57	0	0	0
008 Hurricane Sally	0	0.89	0	0	0.89	0
011 Rutherford Floods	2.86	1.43	4.29	2.86	1.43	1.43
013 Edenville Dam Failure	1.43	0	0	0	0	1.43

Figure 3: Coverage Disparity across ten crisis events, measured by non-target concept scores (percentage of summary items covering tangential topics, e.g., animal welfare). Red cells indicate higher distraction, while green cells indicate low to zero distraction.

6 Discussion

Our findings suggest that while LLMs seem to have achieved fluency, they systemically fail to meet the core requirements of the high-stakes MDDS task. We identify three critical ways in which LLMs fail to meet the demands of the task: (1) the consensus trap in coverage, (2) the safety risks of majority bias and (3) the reliability gap in verification. In this section, we unpack these findings and their implications.

The Coverage Consensus Trap: Breadth without Depth. Our evaluation of the coverage performance of LLMs revealed that while models show relatively high competence in identifying high-level information clusters (Coverage (Cluster)), they struggle to recover the specific and granular details that define the representative item (Coverage (Gold)). The coverage consensus trap refines observations from Huang et al. (2024) and Laban et al. (2024) about coverage into a failure mode that is specific to high-stakes MDDS. In short, LLMs tend to function as "smoothing" narrators. When presented with multi-document and multi-source information, they focus on high-level information that captures the main gist. This behavior, although useful for consensus-based summarization (Fabbri et al., 2019; Zhang et al., 2024a), in crisis response, where specific actionable information are critical (e.g., *where* rescue is needed), is a failure mode.

Annotation Challenge and Human Decision-Making. We hypothesize that the consensus trap reflects the underlying difficulty of the MDDS task formulation and annotation design. During the manual annotation process, we found that identifying unique information clusters is a difficult task, specifically to resolving/grouping similar clusters. For example, as shown in Figure 1 reference summary, reports of "80 people at Palomar College Shelter" overlapped in a way with "Overnight, about 615 people took refuge in 13 Community shelters". Rather than forcing resolution, our annotation decision emphasizes that *specific, potentially conflicting information must be preserved.* The final decision-making should be passed on to the human emergency response officer or disaster manager and the LLM should present all specific variations. In high-stakes settings, the human decision-maker must retain full control in resolving ambiguity.

Coverage Disparity is an Event-Specific Concern. We observed that coverage disparity (non-target concept score) is not uniform but highly event-dependent (see Figure 3). While the average non-target concept score was low at <6% (see Table 2), specific events like the 001 Lilac Wildfire went up to over 18%. This complements prior findings that summarization metrics and models can be sensitive to length and topical confounders (Guo and Vosoughi, 2023; Chhabra et al., 2024; Laban et al., 2024). Error analysis revealed that this is driven by event-specific lexical overlap. In wildfire events, for example, reports of "human impact and welfare" frequently overlap with "animal welfare" (e.g., horses affected or sheltered, dogs needing rescue). LLMs seem to struggle to distinguish these topics due to high lexical overlap with frequent use of words such as evacuation, shelter, and rescue. This means that there is a need for disambiguation and event-specific requirements, where current zero-shot settings fail.

Majority Bias as a Safety Risk. Our frequency–salience analysis extends the literature on bias in summarization and language models—often focused on social or position bias (Anantaprayoon et al., 2024; Steen and Markert, 2024; Olabisi and Agrawal, 2024; Ravaut et al., 2024; Zhou et al., 2025)—to content-selection bias over redundant versus rare information. We identify that majority bias is a prevalent issue in our task. Our finding further confirms the observation of Huang et al. (2024), where they noted that long-context LLMs tend to have better

coverage for more frequent answers. We show that frequency can be a source of inequitable selection. Our metric demonstrates that majority bias is not just a qualitative or theoretical concern but a measurable behavior across models. The consistent negative correlation supports the broader claim that LLMs are not neutral information processors (Gallegos et al., 2024; Liang et al., 2023). Instead, they tend to amplify information that is repeated across different source documents and, consequently, pass over infrequent information.

The Reliability Gap in Verification. While verifiable summaries are a prerequisite for trust, especially in high-stakes settings (dela Cruz et al., 2025), our Citation Score average values between 35.11 and 49.60 indicate that more than half of the generated facts are not properly grounded in source documents, a finding in line with Laban et al.'s (2024) work. Our novel metric, Citation-Frequency correlation adds a frequency-aware perspective that previous benchmarks (Laban et al., 2024; Li et al., 2024) lack. Our results show a reliability gap: negative correlation of Citation Score and frequency of relevant facts for five out of the six LLMs. This negative correlation, although low, means that we observe LLMs tend to misattribute citations exactly when the information is rarer, which is where, in crisis response, our end users would most need to verify the information. This creates a pitfall: models are most reliable verifying frequent/common facts (easier to verify) and least reliable when reporting unique and less frequent facts.

Implications on High-stakes MDDS. Our novel metrics—Comprehensive Coverage Score, Frequency-Salience correlation, and Citation-Frequency correlation—provide useful tools that complements existing metrics rather than replacing them, making explicit the consensus trap, coverage disparity, majority bias, and inequitable grounding that standard evaluation pipelines—including those based on LLM-as-a-judge (Huang et al., 2024; Laban et al., 2024)—cannot reliably expose.

Although we carry out our investigation in crisis response, the same three requirements and metrics are also applicable to other high-stakes MDDS domains where omissions and unverifiable claims can cause harm. For example, in the medical domain, LLM-generated hospital summaries can be evaluated not only on overall quality assessments (Challener et al., 2026), but also in frequency-aware properties. In legal and regulatory settings, case summaries can be evaluated to complement specificity, sensitivity, and clarity metrics (Heddaya et al., 2025). Overall, adopting our high-stakes MDDS requirements as an evaluation template allows practitioners to tailor-fit evaluation to their own domain, while incorporating tests for failure modes that current summarization benchmarks are not designed to detect.

7 Conclusion

We investigate MDDS task in high-stakes settings, where summaries must go beyond consensus by covering all the diverse information, reducing redundancy into unique information clusters, and equitably grounding claims. To support evaluation under these requirements, we construct a new multi-source dataset adapted from CrisisFacts (McCreadie and Buntain, 2023) across ten crisis events, and introduce a manual annotation pipeline that identifies relevant human-impact posts, clusters redundant reports, selects a representative item per cluster, and cites all source documents. We propose a set of novel metrics aligned to task: Comprehensive Coverage Score to capture breadth and depth jointly, Frequency-Salience correlation to quantify majority bias in content selection, and Citation-Frequency correlation to assess grounding equity.

Our analysis of six commonly used LLMs, revealed three critical failure modes. First, the consensus trap shows that models capture high-level information clusters while missing specific actionable details. Second, LLMs systematically exhibit majority bias, prioritizing more frequent facts. Third, there is a verifiability gap among most of the LLMs, except GPT 5-mini, meaning there are unreliable citations precisely when verification is most needed for rare facts. Our proposed evaluation framework offers a practical and interpretable toolkit for diagnosing coverage, majority bias, and grounding equity not only in crisis response but can also be adopted in other high-stakes settings where omissions and weakly grounding of information cause harm.

Limitations

Our study has several limitations. We focus on English-only crisis events from 2017-2020 collected from CrisisFacts (McCreadie and Buntain,

2023); generalizability to other languages, time periods, or crisis types requires validation. Our dataset construction was based on a single expert annotator and validated with a 10% sample based on another expert annotator. Our human annotator (first author) is fluent in English and has relevant experience in crisis informatics and corpus linguistics. The second annotator is fluent in English and has relevant experience in corpus and computational linguistics. The inter-rater reliability scores, measured across each step of the annotation process, suggest that the annotation pipeline is reliable across all three stages, with strongest agreement in relevance annotation and substantial agreement in both clustering and representative selection (Appendix A.2). Representative selection remains the most subjective stage, but the observed macro F1-score indicates substantial alignment once relevance is established.

We note that we restricted our output summaries into seven bullets based on the maximum number clusters in the reference summaries. This was not tested to be an optimal deployment setting. This length constraints could have had possible interaction with our metrics.

Our Frequency-Salience and Citation-Frequency metrics focus on the relevant set (clustered for redundancy) rather than the full document set, limiting our analysis of how **non-relevant** but frequent topics distract models. Our Non-target concept score is based on keyword matching and could be brittle and not capture off-topic concepts that use paraphrasing. This score is introduced as a measure within the maximizing information coverage requirement and can still be improved. Finally, we test off-the-shelf models without exploring prompting strategies or fine-tuning approaches that might mitigate identified biases.

Ethical Considerations

The source dataset used in this paper are derived from publicly available dataset which were collected tweets from X (previously, Twitter), Facebook Posts, Reddit Posts and News items. For more information about the data collection, see McCreadie and Buntain's (2023) paper. The dataset used in this work contains crisis-related content including reports of deaths, injuries, missing persons, and displaced populations. Theselected sample, the constructed manual gold summary dataset constructed and the annotation protocol are available in Dela Cruz et al. (2026). Readers and future users of this dataset may find this content distressing.

Our work aspires to support practitioners in high-stakes scenarios. Our work demonstrates measurable progress of LLMs in the high-stakes MDDS task. We note that current performance of LLMs in the current prompt configuration, they do not achieve the task-specific requirements needed for unsupervised deployment. Therefore, there is a potential risk that readers misinterpret the readiness of the technology for use, and move too quickly to adapt without guarantees of reliability. However, we also caution against the premature abandonment of the use of LLMS for the MDDS task. The diagnosable failure modes of these LLMs in the task suggest that targeted methods for iterative improvement can be explored further with domain expertise.

We recognize that the annotation work for this summarization task carries emotional burden. The documents included posts that contain reports of injuries, casualties, missing persons, and widespread infrastructure damage. Asking annotators to read, group, and select representative posts carries inherent psychological risks. Annotators were informed and warned beforehand about the sensitive nature of the task prior to participation. They were encouraged to take breaks or opt out if the task became overwhelming.

Acknowledgments

This publication is part of the project 'Indeep: Interpreting Deep Learning Models for Text and Sound' with project number NWA.1292.19.399, which is partly financed by the Dutch Research Council (NWO). We thank Wenwei Dong for helping in the annotation work.

References

Gianni Amati, Giuseppe Amodeo, Marco Bianchi, Giuseppe Marcone, Fondazione Ugo Bordoni, Carlo Gaibisso, Giorgio Gambosi, Alessandro Celi, Cesidio Di Nicola, and Michele Flammini. 2011. Fub, iasi-cnr, UNIVAQ at TREC 2011 microblog track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011*, volume 500-296 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Panatchakorn Anantaprayoon, Masahiro Kaneko, and

Naoaki Okazaki. 2024. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6395–6408, Torino, Italia. ELRA and ICCL.

Douglas Challener, Shant Ayanian, Alexander Ryu, John O'Horo, and Heather Heaton. 2026. Quality assessment of artificial intelligence-generated versus human-written hospital summaries evaluating detail, usefulness, and continuity of care. *Journal of Hospital Medicine*, 21(4):375–379.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 1–11, Mexico City, Mexico. Association for Computational Linguistics.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Jane Arleth dela Cruz, Iris Hendrickx, and Martha Larson. 2025. Evaluating large language models for confidence-based check set selection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16249–16265, Vienna, Austria. Association for Computational Linguistics.

Jane Arleth Dela Cruz, Iris Hendrickx, and Martha Larson. 2026. Dataset for Beyond Consensus: Diagnosing Coverage, Majority Bias, and Grounding Equity in High-Stakes Multi-Document Diversity Summarization.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2026. A survey on llm-as-a-judge. *The Innovation*, 7(6).

Xiaobo Guo and Soroush Vosoughi. 2023. Length does matter: Summary length can bias summarization metrics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15869–15879, Singapore. Association for Computational Linguistics.

Mourad Heddaya, Kyle MacMillan, Hongyuan Mei, Chenhao Tan, and Anup Malani. 2025. CaseSumm: A large-scale dataset for long-context summarization from U.S. Supreme Court opinions. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1917–1942, Albuquerque, New Mexico. Association for Computational Linguistics.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. 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 570–593, Mexico City, Mexico. Association for Computational Linguistics.

Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903, Miami, Florida, USA. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi R. Wang, and Jundong Li. 2025. Harnessing large language models for disaster management: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14528–14551, Vienna, Austria. Association for Computational Linguistics.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 493–516, Bangkok, Thailand. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

R. McCreadie and C. Buntain. 2023. Crisisfacts: Building and evaluating crisis timelines. In *Proceedings of the 20th International ISCRAM Conference*, pages 320–339, Omaha, USA. University of Nebraska at Omaha.

Mistral AI Team. 2024. Mistral nemo. `https://mistral.ai/news/mistral-nemo`.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Olubusayo Olabisi and Ameeta Agrawal. 2024. Understanding position bias effects on fairness in social multi-document summarization. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 117–129, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2025. Gpt-5 system card.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Seeberger and Korbinian Riedhammer. 2023. Multi-query focused disaster summarization via instruction-based prompting. In *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14-17, 2023*, volume 1328 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Julius Steen and Katja Markert. 2024. Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5962–5983, Bangkok, Thailand. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, and 1 others. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Yixuan Tang, Yuanyuan Shi, Yiqun Sun, and Anthony Kum Hoe Tung. 2025. Uncovering the bigger picture: Comprehensive event understanding via diverse news retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33939–33957, Suzhou, China. Association for Computational Linguistics.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron C. Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9871–9889, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pages 102351–102390.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTSCORE: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024a. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024b. Fair abstractive summarization of diverse perspectives. 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 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2025. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

A Data and Annotation

A.1 Data

Queries about **Human Impact and Welfare**:

- How many people are affected (evacuated)?
- How many people are in shelters?
- How many people are missing?
- How many people are trapped?
- How many people have been injured?
- How many people have been killed?
- What at-risk groups are being impacted?

Preprocessing. We replaced URLs in the posts with <URL> tokens and ran an initial exact match deduplication of posts. We follow a two-stage retrieval approach consisting of DFreeKlim-based retrieval (Amati et al., 2011) and subsequent MonoT5 neural reranking (Nogueira et al., 2020). This two-stage approach was known to perform well with crisis data settings (Seeberger and Riedhammer, 2023). We selected up to the top 15 posts per query per day. This design was based on manual verification, where beyond this count already contained multiple duplicates.

Then, we further deduplicated the posts using a combination of TF-IDF vectorization and cosine similarity at threshold of 0.85 to detect near-duplicate posts, with the post with the highest MonoT5 relevance was kept.

The overview description of the dataset used in the experiment is in Table 4.

A.2 Manual Gold Summary Annotation

The annotation process followed three steps, as shown in Figure 4:

1. **Relevance Annotation:** Posts were annotated to determine if they belonged to the *Human Impact (Relevant Set)*.
2. **Redundancy Check and Cluster Creation:** We checked the Relevant Set for redundant posts. We grouped posts reporting the same information into unique information clusters.
3. **Representative Post Selection:** We selected the single most informative post in each cluster as the representative post. If the posts are very similar, we select the first post as representative.

Inter-Annotator Agreement. We sampled 10% of the 100-post event-day batches to compute for inter-annotator agreement scores for the three levels of annotation following each annotation step

Event	# of days	# of summaries	# of gold facts (clusters)	# of relevant facts (relevant set)	# of documents
All Events	63	126	630	1,125	11,770
Lilac Wildfire 2017	9	19	79	152	1,740
Cranston Wildfire 2018	6	12	71	154	1,115
Holy Wildfire 2018	6	12	42	69	1,200
Hurricane Florence 2019	14	30	158	287	2,675
Maryland Floods 2018	4	8	39	74	800
Saddleridge Wildfire 2019	4	9	66	123	882
Hurricane Laura 2020	2	5	38	49	450
Hurricane Sally 2020	8	12	66	103	1,144
Rutherford Floods 2020	5	9	41	54	818
Edenville Dam Failure 2020	5	10	30	60	946

Table 4: CrisisFacts Dataset Sample used for our experiments. The final dataset includes 11,770 documents from 63 days of crisis. We investigate on 126 individual summaries.

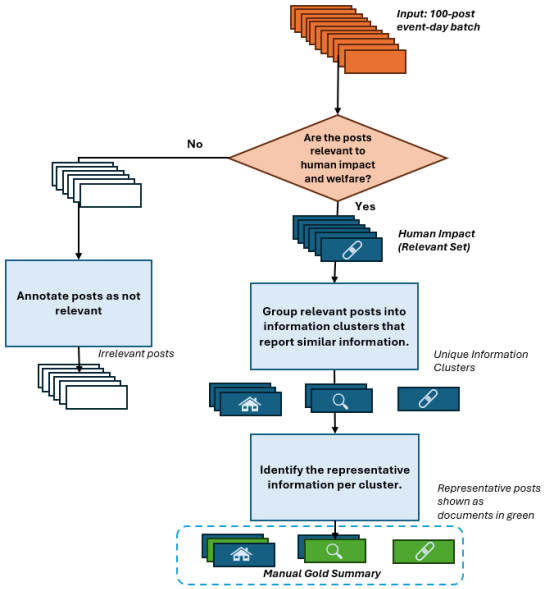


Figure 4: Manual Gold Summary Construction.

from above. The Cohen's kappa (κ) agreement score for the human impact set annotation was 0.84 which indicates an almost perfect agreement. Furthermore, the raw agreement was at 0.97 weighted F1-score on the human impact set annotation, with 0.86 F1-score agreement for the True (Human Impact (Relevant Set)) category.

We evaluated the inter-annotator agreement for redundancy check and cluster creation using Rand Index and Adjusted Rand Index, on the union of the annotators' relevant sets. Since the dataset contains a substantial amount of unique, non-redundant single-item posts (singletons), metrics like pairwise F1-score and Adjusted Rand Index can penalize agreement on singletons. The mean Rand Index per event-day batch was 0.96, while the mean Adjusted Rand Index was 0.70.

We evaluated the inter-annotator agreement for the representative post selection on the intersection of the two annotator's relevant sets using macro F1-score. The event-day batch mean macro F1-score was measured at 0.75, indicating when annotators have agreed on relevance they have substantial alignment in selecting the representative.

B Models

Table 5 contains the information about the versions of the six LLMs we evaluated and analyzed. All five models, except for GPT 5-mini, were set at temperature 0.0 to be deterministic. For GPT 5-mini, we set the reasoning effort to "low".

Model	Type	Source (OpenAI/Huggingface)
GPT 4o-mini	closed	gpt-4o-mini-2024-07-18
GPT 5-mini	closed	gpt-5-mini-2025-08-07
Llama 3.1 - 8B Instruct	open	meta-llama/Llama-3.1-8B-Instruct
Llama 3.2 - 3B Instruct	open	meta-llama/Llama-3.2-3B-Instruct
Mistral NeMo	open	mistralai/Mistral-NeMo-Base-2407
Qwen 3 4B	open	Qwen/Qwen3-4B-FP8

Table 5: Information of evaluated and analyzed LLMs

C Prompt

The prompt used for the MDDS crisis summarization is found in Table 6.

System	You are a concise assistant. Use ONLY the given data. Output only bullet points: one factual sentence + post IDs in brackets. If none found, output: Unanswerable. No thoughts or explanations.
User	You are given **{count}** information posts from the crisis event "**{event}**": **{list of information posts}** Your task: Identify which posts contain relevant information that helps answer the following question: "What critical new developments in how the crisis event, **{event}**, is directly impacting people -- including numbers of casualties, missing persons, injured individuals, people in shelters, at-risk groups, and locations where assistance or rescue is needed -- are relevant and useful for disaster managers?" **Summarize the relevant information** from your identified posts. -- Output rules: 1. You must only output a **bullet list**. 2. For each bullet point: - You must write **one clear and concise sentence** describing a unique piece of relevant information. - You must include one or more reference post IDs in brackets at the end of each sentence. - You may include multiple post IDs if they support the same piece of information. 3. **Do NOT repeat post IDs** across bullets. 4. Output **a maximum of 7 bullet points** (fewer is acceptable). 5. If no relevant posts are found, output a single bullet point: -"Unanswerable" 6. Do not include explanations, reasoning, or text outside the bullet list. -- Example output: - 15 people missing after floods in the western district; rescue operations underway. [News-3-4][Twitter-2-7]

Table 6: Prompt for MDDS crisis summarization for *human impact and welfare* information. We modify the **{count}** for the number of information posts input, **{event}** for the crisis event name, and the **{list of information posts}** include input posts listed chronologically by timestamp, with post IDs and text.