EQUAL: Entity-Enhanced QUery Expansion for EquitAble Crisis Summarization via KnowLedge Graphs

Hajra Klair
Virginia Tech
khajra23@vt.edu

Hoda Eldardiry
Virginia Tech
hdardiry@vt.edu

William A. Ingram
Virginia Tech
waingram@vt.edu

ABSTRACT

Disaster response efforts face persistent challenges in ensuring equitable aid and information access for all affected populations. Marginalized communities—including the elderly, persons with disabilities, people experiencing homelessness, low-income households, non-English speakers, and geographically isolated residents—face heightened risk during disasters and are more likely to experience delays in receiving aid, evacuation support, and critical information (Wilson et al. 2021). We present EQUAL (entity-enhanced query expansion for equitable crisis summarization via knowledge graphs), a work-in-progress framework that combines dual-model entity-enhanced query expansion with an equity-aware GraphRAG (Graph-based Retrieval-Augmented Generation) pipeline. EQUAL constructs crisis knowledge graphs enriched with vulnerability–resource connections and generates summaries through community-level synthesis. Evaluated on 18 real-world disaster events from the TREC CrisisFACTS dataset, EQUAL outperforms all baselines on equity-focused metrics—vulnerable group coverage, intersectional coverage, and statistical parity—and shows marked gains in explicit mentions of vulnerable populations, geographic specificity, and actionable resource information. It also remains competitive on standard semantic quality metrics.

Keywords

crisis informatics, GraphRAG, knowledge graphs, crisis summarization, social vulnerability

INTRODUCTION

Natural disasters disproportionately impact marginalized groups, who face persistent barriers to aid, shelter, and medical care. Richards (2019) examined recent wildfires and hurricanes in the United States and found that low-income, elderly, and disabled populations face disproportionate barriers to emergency support and evacuation. Analyses of federal disaster relief data indicate that aid distribution is shaped by race, ethnicity, and socioeconomic status, often resulting in inequitable recovery outcomes for vulnerable populations (Emrich et al. 2022; Howell and Elliott 2019). These inequalities extend to disaster preparedness infrastructure: people with disabilities routinely encounter inaccessible emergency shelters and warning systems that fail to deliver alerts in accessible formats, compounded by their systematic exclusion from disaster planning processes (Nguyen-Trung et al. 2025).

Comprehensive and inclusive disaster summaries are essential for closing these gaps. They give first responders a clearer picture of evolving needs on the ground and help guide more equitable aid distribution. For researchers and policymakers, they provide representative data to inform disaster resilience planning and social vulnerability analyses. Since the early adoption of social media during disasters (Vieweg et al. 2010), platforms like Twitter, Reddit, and Facebook have become key sources of rapid, location-specific updates from affected communities—a

pattern that persists across recent events (McCreadie and Buntain 2023). Yet the sheer volume and noise in this data make it difficult to extract actionable insights. Manually sifting through thousands of posts per event is not feasible for time-pressed responders. Automated multi-document summarization offers a path forward. To this end, NIST and DARPA established the TREC CrisisFACTS track as a benchmark for evaluating crisis summarization systems (McCreadie and Buntain 2023).

Retrieving relevant information from high-volume, multi-source crisis data is a long-standing challenge in disaster informatics. Sparse retrieval methods like BM25 remain widely used as a first stage, but they rely on exact term matching and struggle with the informal, noisy language common on social media (Lamsal et al. 2024). Neural rerankers and dense retrieval models have improved relevance scoring (Seeberger and Riedhammer 2024), yet the downstream summarization step still optimizes primarily for topical coverage and redundancy reduction—not for equitable representation of the diverse groups affected by a crisis. This leaves a critical gap. When crisis summaries omit or underrepresent vulnerable populations, emergency managers risk allocating resources based on an incomplete picture (C. Zhang et al. 2021).

These equity concerns sit alongside a set of well-documented technical challenges. Annotated datasets are scarce, limiting the use of supervised methods. The information need during an event is unusually complex, often requiring dozens of queries simultaneously. Multi-stream settings produce large, noisy document collections that exceed the input limits of transformer-based models. Target summaries for disaster reporting also tend to be longer than those in typical summarization tasks, which remains challenging for abstractive methods (Seeberger and Riedhammer 2024).

While recent work has made progress on these technical fronts, the question of who is represented in the resulting summaries has received far less attention. In this work, we target that gap by integrating equity considerations into the retrieval and selection pipeline. More specifically, we frame the task as follows. Given a predefined set of M queries, each representing a distinct information need during a crisis, the goal is to retrieve and summarize relevant content from a diverse document collection of N disaster-related documents gathered from news articles, social media, and official reports. Beyond topical relevance, we additionally require that the selected documents capture a wide range of perspectives and local context—especially with regard to underrepresented groups or impacted communities that may be excluded from mainstream coverage.

Our Contributions

(1) We develop EQUAL, a crisis summarization framework that expands generic disaster queries into entity-enhanced queries, constructs an equity-aware knowledge graph that amplifies vulnerability–resource connections, and synthesizes inclusive summaries through GraphRAG-based community detection.

(2) We assemble a suite of inclusivity metrics to evaluate whether a given crisis summary equitably represents all affected populations, particularly those most at risk.

(3) We evaluate EQUAL across 18 real-world disasters from TREC CrisisFACTS, combining established metrics (BERTScore) with inclusivity metrics and a granular coverage analysis across equity and geographic dimensions.

RELATED WORK

Crisis Summarization from Social Media

Social media platforms have become a primary source of real-time information during disasters, generating high-volume, multi-source data streams that can support situational awareness and response coordination. Early crisis informatics pipelines focused on filtering and categorizing Twitter data, followed by extractive summarization to distill actionable content (Nguyen and Rudra 2022). Rudra et al. (2015) combined tweet classification with integer linear programming to maximize topic coverage in crisis summaries, while later work explored temporal streaming summarization for alert generation (Dusart et al. 2021). The CrisisFACTS shared task (McCreadie and Buntain 2023) advanced the field further by enabling evaluation of multi-stream summarization systems that draw from Twitter, Reddit, Facebook, and online news simultaneously.

The retrieval stage underpinning these pipelines has evolved in parallel. Sparse methods like BM25 remain widely used for first-stage retrieval, but they depend on exact term matching and perform poorly on the informal, noisy language typical of social media (Lamsal et al. 2024). Neural rerankers such as MonoT5 improve relevance scoring by capturing semantic similarity (Seeberger and Riedhammer 2024), and dense retrieval models further close the vocabulary mismatch gap. On the generation side, recent work integrates QA-inspired fact extraction using chain-of-thought prompting (Pereira et al. 2023) and retrieval-augmented generation (Colverd et al. 2023). Seeberger and Riedhammer (2024) proposed Crisis2Sum, a retrieve-then-summarize framework combining query-focused

ILP-based selection with LLM-based fact extraction, reporting that extractive methods can be competitive with or outperform abstractive baselines for disaster reporting—a finding that underscores the importance of informativeness over fluency in this domain. GraphRAG[1] (Edge et al. 2024) extends traditional RAG by constructing knowledge graphs from source documents, detecting communities using algorithms such as Louvain (Blondel et al. 2008), and generating community-level summaries before synthesizing a global response. Despite this technical progress, existing pipelines optimize almost exclusively for topical coverage and redundancy reduction. They do not consider whether the resulting summaries equitably represent the diverse populations affected by a crisis—a gap that the following subsection situates within the broader disaster equity literature.

Inclusivity and Social Vulnerability in Disaster Informatics

Research consistently shows that people with disabilities, low-income households, and other marginalized populations are disproportionately affected by disasters and frequently overlooked in mainstream response strategies (Stough and Kang 2015; Aldrich 2012). The Sendai Framework for Disaster Risk Reduction explicitly calls for disability-inclusive planning and the participation of vulnerable populations in all phases of disaster management (Stough and Kang 2015). Institutional efforts have produced frameworks for disability-inclusive risk reduction and accessible communication technologies, yet these advances have not been integrated into automated information processing pipelines.

Quantitative frameworks for measuring vulnerability do exist. The Social Vulnerability Index (SoVI) (Cutter et al. 2003) identifies demographic, socioeconomic, and geographic factors that shape disaster vulnerability, and the CDC's Social Vulnerability Index (Flanagan et al. 2011) operationalizes similar concepts for public health preparedness. While well-established in emergency management practice, neither has been connected to automated retrieval or summarization systems. Current evidence further indicates that existing information systems can perpetuate social inequalities through design limitations (Wilson et al. 2021), reinforcing the need for pipelines that actively identify and prioritize at-risk populations. EQUAL addresses this gap by adapting the GraphRAG architecture for crisis contexts, injecting equity awareness at every stage—from entity extraction through graph construction to community-level summarization.

METHODOLOGY

We provide details of EQUAL, our crisis summarization framework designed for inclusive, query-focused disaster summarization. At a high level, EQUAL works in two stages. First, it enriches generic disaster queries with event-specific entities from the document collection, so that retrieval captures locally relevant information that keyword matching would miss. For instance, during the Lilac Fire, a query for "evacuated" alone would not match posts describing the "Bonsall evacuation" or "Camp Pendleton evacuation"; entity expansion bridges this gap by linking corpus-derived entities to base query terms and retrieving against the enriched query set. Second, EQUAL organizes retrieved documents into a knowledge graph linking entities like locations, shelters, and vulnerable populations based on co-occurrence in the source text. Community detection then groups densely connected entities together, so that a nursing home, its nearby medical facility, and the elderly residents it serves end up in the same cluster. This affords the final summary a structure that preserves who is affected and what resources are near them, instead of collapsing everything into a generic overview.

More specifically, EQUAL operates in two major phases: (1) an entity-enhanced query-based document retrieval phase that expands generic disaster queries through dual-model NER extraction and LLM-based entity-to-query mapping; and (2) a GraphRAG summarization phase that constructs an equity-aware knowledge graph from retrieved documents and generates summaries through community-level synthesis.

Entity-Enhanced Query Expansion

The input to EQUAL consists of a document corpus $\mathcal{D} = \{d_1, \ldots, d_N\}$ of event-related documents and a set of base queries $Q = \{q_1, \ldots, q_M\}$, where N and M denote the number of documents and queries, respectively. Standard keyword matching over Q is insufficient when queries are expressed in broad or compound terms. As noted by Seeberger and Riedhammer (2024), information needs during a disaster event are complex and may require considering dozens of queries simultaneously, while documents in $\mathcal{D}$—particularly social media posts—are short, noisy, and high-volume. For example, when emergency responders search for q_i = "food water transport," they may miss documents in $\mathcal{D}$ that describe resource delivery using event-specific language (e.g., "Oceanside High School supply drop during the Lilac Fire").

[1] github.com/microsoft/graphrag

To address this, EQUAL employs dual-model named entity recognition—combining a spaCy statistical model[2] with a transformer-based BERT NER model—to extract event-specific entities from $\mathcal{D}$. The two models are complementary: spaCy provides broad entity type coverage (persons, organizations, geo-political entities, locations, facilities, events, dates, and quantities), while the BERT model better captures contextual entities in noisy social media text. Since $|\mathcal{D}|$ ranges from 20,000 to over 550,000 documents per event, running dual NER on the full corpus is prohibitively expensive. We therefore apply a stratified priority sample of up to 2,000 documents per event, ranking by source type (news articles first) and document length to maximize entity richness. While this does not guarantee exhaustive coverage, news articles and longer texts disproportionately concentrate the named entities most relevant to query expansion, and additional documents yield diminishing returns on novel entity vocabulary. Both entity sets are merged, deduplicated, and cleaned to form a corpus-level entity set:

$$E_{\mathcal{D}} = \bigcup_{d_j \in \mathcal{D}} \text{NER}(d_j) \tag{1}$$

Following recent work on LLM-based query expansion (Wang et al. 2023), we use an LLM (GPT-5-mini) to map the extracted entities $E_{\mathcal{D}}$ to the base query terms in Q. The model receives up to 200 entities and the full set of query terms, and returns a mapping that associates each q_i with its relevant entities, yielding an enhanced query vocabulary:

$$Q_{q_i}^{\text{enh}} = \{q_i\} \cup \text{LLM-Map}(q_i, E_{\mathcal{D}}) \tag{2}$$

This expansion is context-sensitive to each event (e.g., q_i = "evacuated" $\rightarrow$ $Q_{q_i}^{\text{enh}}$ = {"evacuated", "Bonsall evacuation", "Oceanside evacuation"}).

Retrieval and Ranking

The enhanced queries Q^{enh} are used with BM25 retrieval (Robertson and Zaragoza 2009) to retrieve the top-k documents per event from $\mathcal{D}$, with $k = 500$, following established practice in the CrisisFACTS literature (Seeberger and Riedhammer 2024; McCreadie and Buntain 2023). Each candidate document d_j receives an aggregate relevance score computed as the sum of BM25 scores across all enhanced query strings:

$$\text{Score}(d_j) = \sum_{q \in Q^{\text{enh}}} \text{BM25}(d_j, q) \tag{3}$$

Documents are ranked by $\text{Score}(d_j)$, with deduplication by text content to ensure diversity. We will use $\mathcal{D}_k \subset \mathcal{D}$ to denote the resulting retrieved set.

Equity-Aware Named Entity Recognition

Standard NER pipelines extract general-purpose entities such as persons, locations, and organizations, but lack the capacity to identify mentions of *vulnerable populations* or *disaster-specific resources*—precisely the information most critical for equitable crisis response. To bridge this gap, the retrieved set $\mathcal{D}_k$ is processed through a custom NER pipeline that extends spaCy's entity recognition with an equity-aware vulnerability taxonomy derived from SoVI (Cutter et al. 2003). This taxonomy captures population groups that disaster research consistently identifies as disproportionately affected yet underrepresented in automated information systems.

Specifically, EQUAL defines a set of eight vulnerability categories $\mathcal{V}$ = {`VUL_ELDERLY`, `VUL_DISABLED`, `VUL_CHILDREN`, `VUL_HOMELESS`, `VUL_LOW_INCOME`, `VUL_NON_ENGLISH`, `VUL_MEDICAL`, `VUL_ISOLATED`} and four resource entity types $\mathcal{R}$ = {`RESOURCE_SHELTER`, `RESOURCE_MEDICAL`, `RESOURCE_TRANSPORT`, `RESOURCE_FOOD`}. Together, $\mathcal{V}$ and $\mathcal{R}$ enable downstream graph construction to explicitly model *who* is vulnerable and *what* resources are available to them. Recognition employs compiled regex pattern matching (e.g., "nursing home residents" $\rightarrow$ `VUL_ELDERLY`) alongside contextual enrichment, where standard entities are re-labeled based on surrounding vulnerability keywords within a fixed character window. For instance, in documents from the Lilac Fire event, the `FAC` entity "Del Mar Fairgrounds" appears near the keywords "elderly" and "medical needs"; the contextual enrichment step assigns it the additional labels `VUL_ELDERLY` and `VUL_MEDICAL`, signaling that this facility serves vulnerable populations.

[2] `spacy.io/models/en#en_core_web_lg`

Crisis Knowledge Graph Construction

Entities extracted from $\mathcal{D}_k$ are assembled into a crisis knowledge graph $G = (V, E)$ using NetworkX.[3] Each node $v \in V$ represents a unique entity, deduplicated by deterministic hashing of its normalized text and label; repeated mentions across documents consolidate into a single node with accumulated mention counts and merged equity labels.

Edges in E encode co-occurrence: for each document in $\mathcal{D}_k$, we identify entity pairs (e_i, e_j) appearing within a fixed character window. The base weight is inversely proportional to their character distance, so that adjacent entities receive stronger connections:

$$w_{\text{base}}(e_i, e_j) = 1.0 - \frac{\text{dist}(e_i, e_j)}{\text{window\_size}} \tag{4}$$

where $\text{dist}(e_i, e_j)$ is the character-level distance between the end of e_i and the start of e_j, and window_size is the maximum co-occurrence range.

To ensure that connections between vulnerable populations and available resources survive community detection, EQUAL applies an *equity boost* $\phi = 1.5$ at two pipeline stages: once during relation extraction and once during graph construction. Equation 5 shows the per-stage boost; because it is applied twice, $\mathcal{V}$–$\mathcal{R}$ edges receive a cumulative $\phi^2 = 2.25\times$ amplification:

$$w(e_i, e_j) = w_{\text{base}}(e_i, e_j) \cdot \phi^2, \quad \phi = \begin{cases} 1.5 & \text{if } e_i \in \mathcal{V} \text{ and } e_j \in \mathcal{R} \\ 1.0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{V}$ and $\mathcal{R}$ denote the vulnerability and resource entity sets defined in the previous subsection. The per-stage value of 1.5 was selected empirically to produce a cumulative boost (2.25×) strong enough to keep $\mathcal{V}$–$\mathcal{R}$ edges from being pruned during community detection, yet small enough that non-equity edges still dominate the overall graph topology; systematic tuning of ϕ is left to future work.

Community Detection and Summarization

Following the GraphRAG paradigm (Edge et al. 2024), the Louvain community detection algorithm (Blondel et al. 2008) partitions G into groups of densely connected entities. Intuitively, entities that frequently co-occur across crisis documents—such as a shelter name, its location, and the populations it serves—tend to form a community. Each community is characterized by its dominant entity types and vulnerability labels from $\mathcal{V}$, which together form the community's *themes*. Communities with overlapping themes are then merged into broader clusters through greedy pairwise aggregation, producing a multi-level hierarchy that ranges from fine-grained local topics to coarser event-wide themes.

Each community is summarized by GPT-5-mini[4] via a structured prompt that instructs the model to pay special attention to the vulnerable population categories in $\mathcal{V}$ and to extract specific locations, resources, and status updates. These community-level summaries are then synthesized into a *global event summary* through a second LLM call. Because the equity boost ϕ amplifies $\mathcal{V}$–$\mathcal{R}$ edges during graph construction, vulnerability-related entities are more likely to cluster together in G, helping ensure that information about vulnerable populations is preserved rather than diluted across a generic summary.

EXPERIMENTAL SETUP

Dataset

We use the TREC CrisisFACTS 2022/2023 dataset—a large-scale, multi-source benchmark from NIST and DARPA designed for disaster response research (McCreadie and Buntain 2023). The dataset covers 18 major disasters including wildfires, hurricanes, floods, industrial accidents, and explosions, each spanning 2–15 days. For each event, it aggregates 20,000 to over 550,000 timestamped documents per day from four sources: news articles, keyword-filtered Twitter posts, Reddit posts and comments, and public Facebook pages. Each event features 48–56 daily queries derived from FEMA ICS 209 reports. Table 6 in the appendix provides per-event details.

Seeberger and Riedhammer (2024) observed that the NIST reference summaries, constructed through a pooling-based annotation process, may introduce bias toward participating systems. They found that system rankings differed across NIST, Wikipedia, and ICS-209 reference summaries. We evaluate against both NIST and Wikipedia references.

[3] https://networkx.org

[4] All LLM calls use `temperature`= 1.0 and `max_completion_tokens`= 16,384.

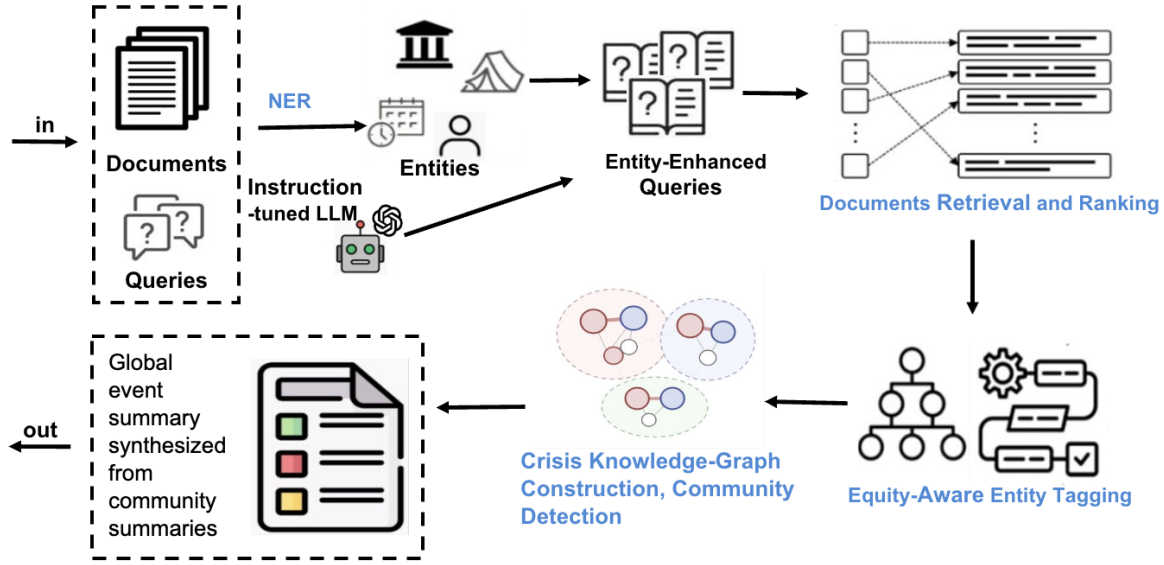

Figure 1. Overview of the EQUAL pipeline. Documents $\mathcal{D}$ and queries Q are provided as input. Dual-model NER extracts entities, which an instruction-tuned LLM maps to base queries to produce entity-enhanced queries Q^{enh}. These are used for BM25-based document retrieval and ranking over $\mathcal{D}$, yielding the retrieved set $\mathcal{D}_k$. Retrieved documents undergo equity-aware entity tagging with vulnerability ($\mathcal{V}$) and resource ($\mathcal{R}$) labels, from which a crisis knowledge graph G is constructed and partitioned into communities. Community-level summaries are synthesized into a global event summary as output.

Baselines

We compare EQUAL against three baselines evaluated under identical conditions (same BERTScore model, same reference summaries):

- **CrisisFACTS Baselines v1 and v2** (McCreadie and Buntain 2023) are extractive systems from the official TREC shared task. Both rank facts by importance and concatenate the top-64 per event, capped at 6,400 characters. Baseline v1 rewards comprehensiveness, measured as the fraction of unique reference facts covered:

$$\text{Comprehensiveness}(S_d) = \frac{\sum_{\{f \in F : M(f,S) \neq \emptyset\}} R(f)}{\sum_{f \in F} R(f)} \tag{6}$$

  where F is the set of all reference facts, $M(f, S)$ is the set of system-items in S that matched fact f, and $R(f)$ is the gain assigned to fact f (set to 1 for CrisisFACTS). Baseline v2 additionally penalizes redundancy by optimizing the ratio of unique facts to total fact matches:

$$\text{RedundancyRatio}(S_d) = \frac{\sum_{\{f \in F : M(f,S) \neq \emptyset\}} R(f)}{\sum_{f \in F} R(f) \cdot |M(f,S)|} \tag{7}$$

- **TREC Abstractive Baseline** follows the evaluator-developed summarization approach: for each event-day, BM25 retrieves relevant sentences per query, GPT-5-mini summarizes each query's facts into 1–3 sentences, the per-query summaries are aggregated and rewritten into a coherent event-day summary, and all event-day summaries are synthesized into a final event-level summary. This baseline uses the same LLM as EQUAL but without entity-enhanced query expansion or the GraphRAG pipeline, isolating the contribution of EQUAL's novel components.

Reference Summaries

Evaluation is conducted against three sets of reference summaries:

- **NIST references** are curated through a pooling-based annotation process in which human assessors identify and label key facts from top-ranked system outputs across all CrisisFACTS participants (McCreadie and Buntain 2023).

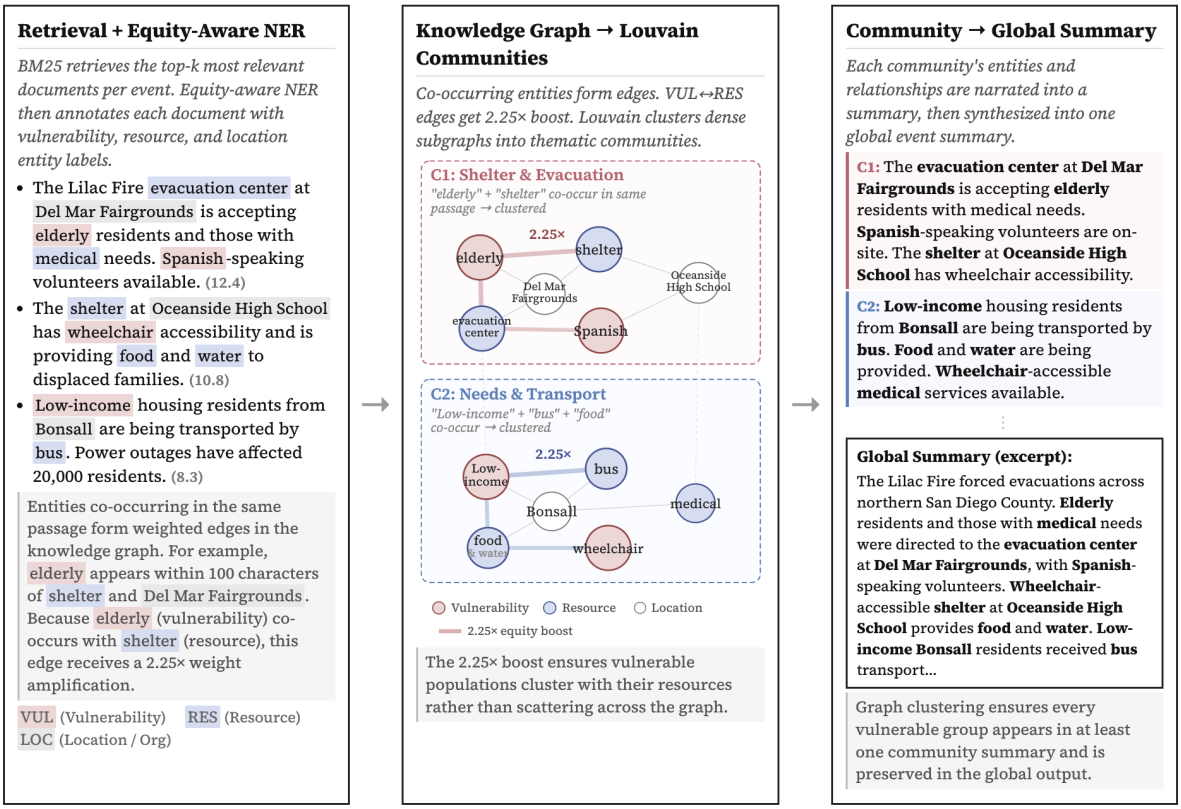

Figure 2. Worked example of EQUAL on Event 001 (Lilac Wildfire) using entities extracted from actual retrieved documents. Left: BM25 retrieves the top-k documents from $\mathcal{D}$, and equity-aware NER annotates each with vulnerability ($\mathcal{V}$), resource ($\mathcal{R}$), and location labels. Center: Co-occurring entities form weighted edges in G; $\mathcal{V}$–$\mathcal{R}$ edges receive a cumulative 2.25× equity boost from ϕ, causing Louvain to cluster vulnerable populations with their associated resources into thematic communities (e.g., C1: *Shelter & Evacuation*, C2: *Needs & Transport*). Right: Each community is narrated into a summary, then synthesized into a global event summary that preserves information about every vulnerable group.

- **Wikipedia references** are drawn from high-level event descriptions on Wikipedia, providing an independent summary perspective not influenced by any participating system.
- **LLM-generated references** are generated by first retrieving the top-k documents from $\mathcal{D}$ using the base queries Q via BM25, then prompting Claude Sonnet 4.5 with these retrieved documents as context alongside the corresponding queries to produce abstractive summaries that serve as an additional LLM-based reference point.

Following the multi-reference evaluation protocol recommended by Seeberger and Riedhammer (2024), we report results against all three reference types to account for potential biases introduced by any single reference source.

EVALUATION METRICS

We evaluate EQUAL along two dimensions—*semantic quality* and *equity*—recognizing that a crisis summary can be semantically adequate yet fail to represent vulnerable populations. Standard summarization metrics assess whether generated text captures the meaning of reference documents, but they are blind to *whose* needs are represented. A summary that thoroughly covers infrastructure damage while omitting all mentions of elderly evacuees or non-English speakers would score well on semantic similarity yet be inadequate for equitable disaster response. We therefore evaluate both dimensions explicitly.

Semantic Quality Metrics

For comparability with prior CrisisFACTS systems, we report **BERTScore** F1 (T. Zhang et al. 2020) using DeBERTa-large-mnli[5], the standard metric in this evaluation track (McCreadie and Buntain 2023; Seeberger

[5] `huggingface.co/microsoft/deberta-large-mnli`

and Riedhammer 2024). BERTScore computes token-level cosine similarities between candidate and reference sentences using contextual embeddings, capturing semantic overlap beyond exact n-gram matching. Including BERTScore allows us to demonstrate that EQUAL maintains competitive semantic quality while pursuing equity objectives. We note that Seeberger and Riedhammer (2024) observed that 83% of NIST reference summaries exceed DeBERTa's 512-token context limit, which may affect discriminative power for longer references.

Equity and Inclusivity Metrics

To directly measure whether summaries represent all affected populations, we supplement BERTScore with five equity and diversity metrics. These metrics are motivated by the disaster equity literature: vulnerability indices like SoVI (Cutter et al. 2003) identify specific demographic groups that face disproportionate risk, yet no existing summarization metric checks whether these groups appear in system output. Our equity metrics bridge this gap by operationalizing the question: *does the summary mention the populations most in need?*

Vulnerable Group Coverage (VGC) measures the fraction of predefined vulnerability categories $\mathcal{G}$ mentioned in a summary S:

$$\text{VGC}(S) = \frac{|\{g \in \mathcal{G} : \exists\, m \in S,\ \text{Match}(m, g)\}|}{|\mathcal{G}|} \tag{8}$$

where $\mathcal{G}$ is the set of vulnerability categories (Table 10) and $\text{Match}(m, g)$ indicates a keyword match. VGC answers the most basic equity question: *how many vulnerable groups are mentioned at all?*

Intersectional Coverage Score (ICS) goes further by capturing co-occurring vulnerability dimensions within the same sentence—reflecting that real-world crises produce intersectional needs (e.g., elderly residents who are also non-English speakers):

$$\text{ICS}(S) = \frac{|\{(g_i, g_j) \in \mathcal{G}^2 : i \neq j,\ \text{CoOccur}(g_i, g_j, S)\}|}{|\mathcal{G}| \cdot (|\mathcal{G}| - 1)/2} \tag{9}$$

where $\text{CoOccur}(g_i, g_j, S)$ indicates that both groups g_i and g_j are mentioned within the same sentence.

Subtopic Recall (SR) measures the fraction of vulnerability subtopics present in the reference that are also mentioned in the generated summary. Subtopics correspond to the seven SoVI groups; a group counts as "present" if any of its lexical indicators appear in the text. See appendix for more details.

$$\text{SR}(S, R) = \frac{|\mathcal{G}(S) \cap \mathcal{G}(R)|}{|\mathcal{G}(R)|} \tag{10}$$

where S is the generated summary, R is the reference summary, and $\mathcal{G}(\cdot)$ denotes the set of SoVI groups detected in a text. SR = 1 when the summary covers every vulnerability group that the reference covers.

α-nDCG (Clarke et al. 2008) extends standard nDCG with a redundancy penalty to reward both relevance and diversity. The summary S is split into sentences and treated as a ranked list; each sentence's gain is determined by which SoVI subtopics it covers, with diminishing returns for subtopics already covered by earlier sentences. The parameter $\alpha \in [0, 1]$ controls this penalty: higher α penalises repetition more heavily. We use $\alpha = 0.5$, and normalise by an ideal ranking in which each reference subtopic appears exactly once in the top-ranked positions.

Statistical Parity (SP) (Dwork et al. 2012) measures whether the summary preserves the reference's relative attention to each vulnerability group. For each SoVI group $g \in \mathcal{G}(R)$, let $n_S(g)$ and $n_R(g)$ denote the number of keyword mentions of group g in the summary S and reference R respectively. The representation rate $r(g) = n_S(g) \,/\, n_R(g)$ captures how much the summary amplifies or attenuates each group relative to the reference:

$$\text{SP}(S, R) = \frac{\min_{g \in \mathcal{G}(R)} r(g)}{\max_{g \in \mathcal{G}(R)} r(g)} \tag{11}$$

SP = 1 indicates that all groups are amplified or attenuated at the same rate; lower values indicate disproportionate over- or under-representation. A summary satisfies the 80% rule when $\text{SP} \geq 0.8$.

Disaggregated Representation Analysis

Beyond aggregate metrics, we conduct a disaggregated representation analysis across all 18 events along two axes: explicit recognition of vulnerable groups and geographic diversity including rural representation. This analysis complements the automatic metrics above, which can be influenced by summary length, extractive vs. abstractive bias in reference construction, and tokenizer truncation.

RESULTS

BERTScore

Table 1. BERTScore F1 (DeBERTa-large-mnli), averaged across 18 events. Best per column in bold.

System	*vs Wiki*	*vs NIST*	*vs Claude*
Baseline v1 (extractive)	0.4356	0.4983	0.4501
Baseline v2 (extractive)	0.4527	**0.5090**	0.4586
TREC Abstractive Baseline	**0.5079**	0.4897	0.4930
EQUAL (ours)	0.5047	0.5050	**0.5007**

Table 2. Equity and inclusivity metrics. Vulnerable Group Coverage (VGC) and Intersectional Coverage Score (ICS) are standalone metrics. Subtopic Recall (SR), α-nDCG, and Statistical Parity (SP) are computed against each reference type. Best per column in bold.

	Standalone		*vs Wiki*			*vs NIST*			*vs Claude*		
System	VGC	ICS	SR	α-nDCG	SP	SR	α-nDCG	SP	SR	α-nDCG	SP
Baseline v1 (extractive)	0.257	0.000	0.875	0.780	0.875	0.264	0.124	0.000	0.273	0.115	0.056
Baseline v2 (extractive)	0.285	0.000	0.953	0.810	0.875	0.293	0.178	0.000	0.346	0.167	0.056
TREC Abstractive Baseline	0.368	0.049	0.844	0.807	0.896	0.374	0.208	0.005	0.432	0.223	0.063
EQUAL (ours)	**0.792**	**0.438**	**1.000**	**0.842**	**0.934**	**0.717**	**0.422**	**0.022**	**0.652**	**0.352**	**0.084**

Table 1 reports BERTScore F1 using DeBERTa-large-mnli, the standard evaluation metric for CrisisFACTS systems. All four systems cluster within a narrow range (0.49–0.51 on NIST, 0.44–0.51 on Wiki), with system rankings shifting across reference types—an instability consistent with the findings of Seeberger and Riedhammer (2024). EQUAL achieves competitive BERTScore across all three reference types, demonstrating that the equity-oriented components of our pipeline do not degrade semantic quality relative to both extractive and abstractive baselines. This is an important result: it establishes that EQUAL's substantial fairness improvements (Table 2) are achieved without sacrificing the summary coherence expected by the summarization community.

We note, however, that BERTScore measures semantic similarity to a reference at the embedding level and does not intrinsically capture the inclusivity or severity of the information present in a summary. A summary that omits vulnerable populations entirely may score comparably to one that covers them, provided both align with the reference surface. The DeBERTa 512-token truncation further limits its discriminative power, as 83% of NIST references exceed this limit (Seeberger and Riedhammer 2024). We therefore complement BERTScore with fairness and diversity metrics that directly assess whether summaries serve the needs of all affected populations.

Equity Coverage

Table 2 reports equity and inclusivity metrics. EQUAL consistently outperforms all baselines across both standalone and reference-based measures. On standalone metrics, EQUAL reaches a VGC of 0.792, compared to 0.368 for the TREC abstractive baseline and 0.257–0.285 for the extractive baselines. Its ICS of 0.438 contrasts with 0.049 for the TREC abstractive baseline and zero for both extractive systems. The TREC abstractive baseline uses the same LLM (GPT-5-mini) yet achieves a low ICS, indicating that LLM summarization alone does not produce intersectional coverage without equity-aware graph construction. Against Wikipedia references, EQUAL achieves maximum Subtopic Recall (1.000), Statistical Parity of 0.934, and the highest α-nDCG across all reference types.

These results also reveal a trade-off between semantic similarity and equity coverage. The baselines achieve higher BERTScore (Table 1) but lower inclusivity scores, while EQUAL achieves a BERTScore of 0.505 (vs Wiki) with substantially higher equity coverage. This reflects the GraphRAG pipeline's emphasis on preserving vulnerability–resource relationships at a modest cost to surface-level semantic similarity.

Disaggregated Representation Analysis

We conducted a disaggregated representation analysis across all 18 event summaries, counting geographic diversity markers, vulnerable group mentions, and medical resource specificity. Findings were independently corroborated by LLM-based structured extraction using GPT-5-mini, which confirmed the same directional results across all reported metrics.

Geographic Diversity

Table 3. Geographic diversity across 18 events (keyword counts).

Mention Type	*Baseline v1*	*Baseline v2*	*TREC Abstractive*	*EQUAL*
Rural/Remote Areas	33	33	4	**124**
Neighborhood-Level Detail	1	2	1	**40**

EQUAL produces substantially more rural/remote area mentions and neighborhood-level detail than all baselines (Table 3). Extractive baselines surface more total location strings overall due to tweet repetition, but lack depth in under-served geographic areas.

Vulnerable Group Representation

Table 4. Vulnerable group mentions across 18 events (keyword counts).

Category	*Baseline v1*	*Baseline v2*	*TREC Abstractive*	*EQUAL*
Elderly	0	0	3	**47**
Disabled/Disability	0	0	5	**52**
Homeless	1	3	0	**8**
Low-Income	4	6	7	**17**
Medically Fragile	8	11	1	**23**
Mental Health	0	1	1	**3**

EQUAL leads in every vulnerable group category (Table 4). Both extractive baselines produce zero mentions of elderly, disabled, and mental health populations across all 18 events. The TREC abstractive baseline, despite using the same LLM, achieves only marginal coverage—indicating that equity-aware graph construction, not the language model, drives vulnerable group representation.

Ablation Study

To isolate the contribution of each pipeline component, we evaluate three ablation variants (Table 5):

- **Ablation 1** (Basic Queries + Direct LLM): Uses base event queries without entity enhancement, and summarizes directly via the LLM without the GraphRAG pipeline.
- **Ablation 2** (Enhanced Queries + Direct LLM): Adds entity-enhanced query expansion, but still bypasses the GraphRAG pipeline.
- **Ablation 3** (Basic Queries + GraphRAG): Uses base event queries but routes retrieval through the full GraphRAG pipeline (graph construction, community detection, and equity-aware summarization).

Table 5. Ablation study results. Each row toggles entity-enhanced query expansion and the GraphRAG pipeline on (✓) or off (×). Subtopic Recall is computed against Wikipedia references.

System	*Entity Expansion*	*Graph RAG*	*Group Coverage*	*Intersectional Coverage*	*Subtopic Recall*
Ablation 1 (Basic Queries + Direct LLM)	×	×	0.576	0.173	1.000
Ablation 2 (Enhanced Queries + Direct LLM)	✓	×	0.528	0.105	0.969
Ablation 3 (Basic Queries + GraphRAG Pipeline)	×	✓	0.806	0.432	0.984
EQUAL (ours)	✓	✓	0.792	0.438	1.000

The GraphRAG pipeline accounts for the largest gains. Adding it while holding queries constant—Ablation 1 vs Ablation 3, and Ablation 2 vs EQUAL—increases VGC by 0.23–0.26 and ICS by 0.26–0.33. This indicates that community detection and equity-aware graph construction are the primary mechanisms through which vulnerable populations surface in the output.

Entity enhancement plays a different role. Comparing Ablation 1 vs Ablation 2 and Ablation 3 vs EQUAL, it has limited effect on equity metrics but contributes to improved BERTScore (Table 1), suggesting it primarily improves retrieval quality rather than equity coverage.

EQUAL and Ablation 3 reach similar VGC (0.792 vs 0.806), but EQUAL achieves slightly higher ICS (0.438 vs 0.432). This suggests that entity-enhanced retrieval, when combined with GraphRAG, improves capture of co-occurring vulnerability dimensions (e.g., elderly residents with mobility limitations). EQUAL also achieves full Subtopic Recall (1.000) against Wikipedia references, indicating that combining both components preserves topical coverage.

DISCUSSION AND LIMITATIONS

EQUAL demonstrates that equity-aware graph construction can meaningfully change what appears in a crisis summary. Notably, the same LLM without the GraphRAG pipeline produces summaries with markedly lower coverage of vulnerable populations. The key mechanism is the equity boost applied during graph construction, which increases the likelihood that vulnerability–resource relationships survive community detection and appear in the final output. In effect, the pipeline encodes a structural prior: that connections between at-risk groups and available resources deserve amplification.

This design reflects a deliberate stance on what crisis summaries should prioritize. Existing crisis summarization systems—including those evaluated in the CrisisFACTS shared task (McCreadie and Buntain 2023)—optimize for topical coverage and redundancy reduction, but are agnostic to *whose* needs are represented. Crisis summaries produced by existing pipelines routinely cover infrastructure damage and evacuation logistics while omitting mentions of elderly evacuees, non-English speakers, or mobility-impaired residents—information that emergency managers need for equitable resource allocation. EQUAL addresses this by making equity a first-class objective throughout the pipeline—from entity extraction through graph construction to community-level summarization.

Practical Implications

Equity-aware summaries can directly inform how resources are allocated in the early hours of a disaster. Emergency managers working with limited personnel, supplies, and transport need to know which vulnerable groups are affected and where. EQUAL's output surfaces exactly this. Consider Hurricane Florence (Event 004), where over 375,000 documents span 15 days across the Carolinas. EQUAL's summaries flagged non-English speaking communities in need of bilingual evacuation guidance and identified low-income households lacking transport to reach designated shelters. Standard baselines produced zero mentions of either group. Without such coverage, responders plan around an incomplete picture. We note that integrating EQUAL into operational workflows still requires validation through user studies with practitioners. However, the substantial gap in vulnerable group coverage between EQUAL and current baselines suggests that even an offline tool could meaningfully support post-disaster review, operational debriefs, and resource planning for future events.

Contextual Nature of Vulnerability

EQUAL operationalizes vulnerability through a fixed set of SoVI-derived categories (Cutter et al. 2003). These capture well-documented demographic risk factors, but vulnerability in practice is relational and context-dependent (Wisner et al. 2004). Who is most at risk shifts with the hazard, the timing, and the social fabric of the affected area. For example, strong community ties or informal support networks can buffer groups that static indices would classify as high-risk. The current taxonomy provides a principled starting point for ensuring traditionally marginalized groups are not overlooked. Incorporating real-time resilience indicators into the vulnerability model is an important consideration, and one that could substantially improve the framework's sensitivity to local context.

Transferability

The CrisisFACTS dataset is predominantly U.S.-based (17 of 18 events). The pipeline's core components, including entity extraction, graph construction, and community detection, are not tied to any particular geography or language, but the vulnerability taxonomy and lexical indicators would need adaptation for non-U.S. contexts. The SoVI-based taxonomy reflects U.S. demographic categories and would need to be replaced with locally appropriate frameworks elsewhere. The lexical indicators are English-only and can be adapted for other languages and cultural contexts.

Limitations

Several additional limitations bound these findings. The equity metrics we introduce offer a complementary evaluation lens, though they rely on keyword-based detection that may miss implicit references or overcount in some cases. EQUAL uses a multi-stage pipeline that may introduce latency unsuited for real-time response, though at approximately 14 minutes per event, this may be an acceptable cost for ensuring vulnerable populations are not omitted. The equity boost parameter has not been systematically tuned, and handling of implicit group references and local jargon remains limited.

CONCLUSION AND FUTURE WORK

We presented EQUAL, a framework that integrates equity awareness into crisis summarization. EQUAL enriches retrieval through entity-enhanced query expansion, identifies at-risk populations via vulnerability-focused NER, and preserves their representation in the final output through GraphRAG-based generation. We evaluated EQUAL across 18 TREC CrisisFACTS disasters. It achieves competitive BERTScore while substantially improving vulnerable group coverage and intersectional representation. The ablation study confirms that the GraphRAG pipeline is the primary driver of equity improvements, while entity enhancement contributes to retrieval quality.

We acknowledge that EQUAL's static vulnerability taxonomy does not capture the context-dependent nature of vulnerability, and that adaptation would be needed for non-U.S. contexts. Several directions remain open. The keyword-based vulnerability detection could be replaced with learned classifiers to improve recall, and adaptive vulnerability models could incorporate real-time resilience indicators alongside static demographic factors. A logical next step is to integrate EQUAL into emergency management workflows and evaluate its impact on resource allocation decisions through user studies with practitioners.

ACKNOWLEDGMENTS

Generative AI Disclosure: EQUAL uses OpenAI's GPT-5-mini for query enhancement and final summarization (the content to summarize is provided). The LLM is used solely for text generation within structured prompts; all evaluation, entity recognition, graph construction, and metric computation are performed by deterministic code. The paper text was drafted with assistance from Claude (Anthropic) for language editing.

REFERENCES

Aldrich, D. P. (2012). *Building Resilience: Social Capital in Post-Disaster Recovery*. Chicago, IL: University of Chicago Press.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008.

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). "Novelty and Diversity in Information Retrieval Evaluation". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 659–666.

Colverd, G., Darm, P., Silverberg, L., and Kasmanoff, N. (2023). *FloodBrain: Flood Disaster Reporting by Web-based Retrieval Augmented Generation with an LLM*. arXiv: 2311.02597.

Crawford, K. (2017). "The Trouble with Bias". In: *Keynote at the 31st Conference on Neural Information Processing Systems (NIPS)*.

Crenshaw, K. (1989). "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". In: *University of Chicago Legal Forum* 1989.1, pp. 139–167.

Cutter, S. L., Boruff, B. J., and Shirley, W. L. (2003). "Social Vulnerability to Environmental Hazards". In: *Social Science Quarterly* 84.2, pp. 242–261.

Dusart, A., Pinel-Sauvagnat, K., and Hubert, G. (2021). "ISSumSet: A Tweet Summarization Dataset Hidden in a TREC Track". In: *Proceedings of the 36th ACM/SIGAPP Symposium on Applied Computing (SAC)*. ACM, pp. 665–671.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*. ACM, pp. 214–226.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. (2024). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. arXiv: 2404.16130.

Emrich, C. T., Aksha, S. K., and Zhou, Y. (2022). "Assessing distributive inequities in FEMA's disaster recovery assistance fund allocation". In: *International Journal of Disaster Risk Reduction* 74, p. 102855.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., and Lewis, B. (2011). "A Social Vulnerability Index for Disaster Management". In: *Journal of Homeland Security and Emergency Management* 8.1.

Howell, J. and Elliott, J. R. (2019). "Damages Done: The Longitudinal Impacts of Natural Hazards on Wealth Inequality in the United States". In: *Social Problems* 66.3, pp. 448–467.

Lamsal, R., Read, M. R., and Karunasekera, S. (2024). "Semantically Enriched Cross-Lingual Sentence Embeddings for Crisis-related Social Media Texts". In: *Proceedings of the 21st International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Münster, Germany.

McCreadie, R. and Buntain, C. (2023). "CrisisFACTS: Building and Evaluating Crisis Timelines". In: *Proceedings of the 20th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pp. 320–339.

Nguyen, T. H. and Rudra, K. (2022). "Rationale Aware Contrastive Learning Based Approach to Classify and Summarize Crisis-Related Microblogs". In: *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, pp. 1552–1562.

Nguyen-Trung, K., Trinh, T. T. T., Nguyen, P. A., Cong-Lem, N., Do, T. H., Le, T. D., Nguyen, H. G., and Simon, M. (2025). "Vulnerabilities of people with different types of disabilities in disasters: a rapid evidence review and qualitative research". In: *Disasters* 49.3, e12686.

Pereira, J., Fidalgo, R., Lotufo, R., and Nogueira, R. (2023). "Using Neural Reranking and GPT-3 for Social Media Disaster Content Summarization". In: *Proceedings of the 31st Text Retrieval Conference (TREC 2022)*. NIST.

Richards, R. (2019). "After the Fire: Vulnerable Communities Respond and Rebuild". In: *Center for American Progress*.

Robertson, S. and Zaragoza, H. (2009). "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3.4, pp. 333–389.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., and Ghosh, S. (2015). "Extracting Situational Information from Microblogs during Disaster Events: A Classification-Summarization Approach". In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, pp. 583–592.

Seeberger, P. and Riedhammer, K. (2024). "Multi-Query Focused Disaster Summarization via Instruction-Based Prompting". In: *Proceedings of the 32nd Text Retrieval Conference (TREC 2023)*. NIST.

Stough, L. M. and Kang, D. (2015). "The Sendai Framework for Disaster Risk Reduction and Persons with Disabilities". In: *International Journal of Disaster Risk Science* 6.2, pp. 140–149.

Takahashi, B., Tandoc, E. C., and Carmichael, C. (2015). "Communicating on Twitter during a Disaster: An Analysis of Tweets during Typhoon Haiyan in the Philippines". In: *Computers in Human Behavior* 50, pp. 392–398.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 1079–1088.

Wang, L., Yang, N., and Wei, F. (2023). "Query2doc: Query Expansion with Large Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 9414–9423.

Wilson, B., Tate, E., and Emrich, C. T. (2021). "Flood Recovery Outcomes and Disaster Assistance Barriers for Vulnerable Populations". In: *Frontiers in Water* 3, p. 752307.

Wisner, B., Blaikie, P., Cannon, T., and Davis, I. (2004). *At Risk: Natural Hazards, People's Vulnerability and Disasters*. 2nd. London: Routledge.

Zhai, C., Cohen, W. W., and Lafferty, J. (2003). "Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 10–17.

Zhang, C., Yang, Y., and Mostafavi, A. (2021). "Revealing Unfairness in social media contributors' attention to vulnerable urban areas during disasters". In: *International Journal of Disaster Risk Reduction* 58, p. 102160.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). "BERTScore: Evaluating Text Generation with BERT". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

APPENDIX

Dataset Details

Table 6. Details of the CrisisFACTS datasets (2023)

ID	*Event*	*Type*	*Days*	*Tweets*	*Reddit*	*News*	*Facebook*
001	Lilac Wildfire 2017	Wildfire	9	41,346	1,738	2,494	5,437
002	Cranston Wildfire 2018	Wildfire	6	22,974	231	1,967	5,386
003	Holy Wildfire 2018	Wildfire	7	23,528	459	1,495	7,016
004	Hurricane Florence 2018	Hurricane	15	41,187	120,776	18,323	196,281
005	Maryland Flood 2018	Flood	4	33,584	2,006	2,008	4,148
006	Saddleridge Wildfire 2019	Wildfire	4	31,969	244	2,267	3,869
007	Hurricane Laura 2020	Hurricane	2	36,120	10,035	6,406	9,048
008	Hurricane Sally 2020	Hurricane	8	40,695	11,825	15,112	48,492
009	Beirut Explosion 2020	Accident	7	94,892	3,257	1,163	368,866
010	Houston Explosion 2020	Accident	7	58,370	5,704	2,175	6,281
011	Rutherford TN Floods 2020	Flood	5	11,019	475	268	9,116
012	TN Derecho 2020	Storm	7	49,247	1,496	15,425	13,521
013	Edenville Dam Failure 2020	Accident	7	16,527	2,339	961	8,358
014	Hurricane Dorian 2019	Hurricane	7	86,915	91,173	7,507	370,644
015	Kincade Wildfire 2019	Wildfire	7	91,548	10,174	339	35,011
016	Easter Tornado 2020	Tornado	5	91,812	5,070	750	34,343
017	Tornado Apr 2020	Tornado	6	99,575	1,233	217	19,878
018	Tornado Mar 2020	Tornado	6	95,221	16,911	641	87,242

Prompt Details

We detail the three LLM prompts used in EQUAL. All prompts use GPT-5-mini with `temperature`=1.0 and `max_completion_tokens`=16384.

Table 7. Entity-to-query mapping prompt (Stage 1). Sent to GPT-5-mini for query expansion.

System: You are a crisis information analyst. Return only valid JSON.
You are given a crisis event and two lists: (1) A list of ENTITIES extracted from documents about this event. (2) A list of QUERY TERMS representing information needs during the crisis. Crisis Event: {event_title}. Description: {event_description}. Extracted Entities: {entities, max 200}. Query Terms: {query terms as JSON}. For EACH query term, select the entities that are relevant to it, and also generate 3–5 expanded query phrases that combine the query term with relevant entities. For example, if the query term is "evacuated" and the entities include "Bonsall", "Oceanside", "Camp Pendleton", return expanded phrases: ["Bonsall evacuation", "Oceanside evacuation", "Camp Pendleton evacuation"]. Return as JSON where each key is a query term and the value is a list of the original term followed by expanded phrases. Only include terms with relevant entities. Return ONLY valid JSON.

Table 8. Community summary prompt (Stage 6a). Sent as `system` message.

You are a crisis information analyst. Summarize the following entities and their relationships from a crisis event. Focus on actionable information for emergency responders and affected populations. IMPORTANT: Pay special attention to mentions of vulnerable populations including: Elderly/seniors, People with disabilities, Children/families, Low-income communities, Non-English speakers, People with medical needs, Homeless/unhoused individuals, Isolated/remote populations. For each community, extract: (1) Key locations and facilities, (2) Resources available (shelters, food, medical), (3) Vulnerabilities and special needs, (4) Critical status updates. Be concise but comprehensive. Include specific names, numbers, and locations.
User message: Community themes: {themes}. Entities in this community: {entity descriptions, max 50}. Summarize this community's information.

Table 9. Global summary prompt (Stage 6b). Synthesizes community summaries into a final event summary.

You are creating a comprehensive crisis summary. Synthesize the following community summaries into a coherent overview of the crisis event.
REQUIREMENTS: (1) Cover all major aspects: damages, casualties, evacuations, resources, recovery. (2) Explicitly mention impacts on and resources for vulnerable populations. (3) Include geographic coverage (urban, rural, specific neighborhoods). (4) Note any information gaps or underrepresented areas.
Event: {event_title}
Community Summaries: {summaries, max 15×500 chars}
Create a comprehensive summary (500–800 words).

Equity Lexicon

EQUAL defines eight vulnerable population categories, informed by the Social Vulnerability Index (SoVI; Cutter et al. (2003)) and emergency management literature. Each category is operationalised through a curated set of lexical indicators used for entity extraction, equity labeling, and metric computation. Table 10 lists the categories and their indicator terms.

Table 10. Equity-aware entity taxonomy derived from SoVI. Each category is detected via compiled regex patterns over the listed keywords.

Category	*Label*	*Keywords (subset)*
Elderly / Seniors	`VUL_ELDERLY`	elderly, senior(s), older, nursing home, retirement, assisted living, aged, retirees
People with Disabilities	`VUL_DISABLED`	disabled, disability, wheelchair, mobility, special needs, blind, deaf, accessible, ADA
Children / Families	`VUL_CHILDREN`	children, kids, minors, school, daycare, pediatric, infant, toddler, youth
Homeless / Unhoused	`VUL_HOMELESS`	homeless, unhoused, shelter, housing insecure, transient, encampment
Low-Income Communities	`VUL_LOW_INCOME`	low-income, poverty, affordable housing, public housing, Section 8, food bank, welfare, unemployed
Non-English Speakers	`VUL_NON_ENGLISH`	Spanish, translation, interpreter, non-English, bilingual, multilingual, language
Medical Needs	`VUL_MEDICAL`	dialysis, oxygen, medication, chronic, medical equipment, health condition, prescription, insulin
Isolated / Remote	`VUL_ISOLATED`	isolated, remote, cut off, stranded, unreachable, rural, no cell service
Shelter	`RESOURCE_SHELTER`	shelter, evacuation center, red cross, community center
Medical	`RESOURCE_MEDICAL`	hospital, clinic, ambulance, EMT, triage, first aid
Food	`RESOURCE_FOOD`	food, water, supplies, distribution, meals, MRE
Transport	`RESOURCE_TRANSPORT`	bus, transportation, evacuation bus, mobility assistance

Intersectionality Pairs

Compound vulnerabilities arise when individuals belong to multiple at-risk groups simultaneously (e.g., an elderly person with a disability in a rural area). EQUAL tracks nine predefined intersectional pairs, selected for their documented prevalence in crisis settings. A pair is considered *covered* in a summary when both constituent categories co-occur within the same sentence.

Table 11. Predefined intersectional vulnerability pairs.

#	Pair
1	Elderly × Disabled
2	Elderly × Isolated / Remote
3	Elderly × Medical Needs
4	Low-Income × Children / Families
5	Homeless × Medical Needs
6	Non-English Speakers × Low-Income
7	Disabled × Isolated / Remote
8	Rural Location × Isolated / Remote
9	Rural Location × Elderly

Equity Metrics

Vulnerable Group Coverage (VGC). VGC operationalizes the principle of *representational coverage* from fairness literature: a system's output should acknowledge the existence of all relevant protected groups (Crawford 2017). It is computed as the fraction of predefined vulnerability categories (Table 10) for which at least one lexical indicator appears in the summary. The metric is reference-free and ranges from 0 to 1.

Intersectional Coverage Score (ICS). ICS draws on *intersectionality theory* (Crenshaw 1989), which holds that individuals at the intersection of multiple marginalized identities face compounded disadvantages not captured by examining each axis of vulnerability in isolation. ICS measures how many predefined vulnerability co-occurrence pairs (Table 11) are attested within at least one sentence of the summary. Sentence-level co-occurrence is used as a proxy for the summary explicitly linking two groups in the same context.

Geographic Equity Index (GEI). GEI addresses the well-documented *urban bias* in both media reporting and crisis informatics (Takahashi et al. 2015), whereby rural, remote, and neighbourhood-level information is systematically under-represented relative to metropolitan areas. GEI is the ratio of non-urban geographic mentions (rural and neighbourhood indicators) to total geographic mentions. A value of 0.5 indicates balanced coverage; values below 0.5 indicate urban over-representation. When no geographic terms are detected, GEI defaults to 0.5 (no bias assumed).

Subtopic Recall. Introduced by Zhai et al. (2003) for evaluating search result diversity, Subtopic Recall measures the proportion of known subtopics covered by a system's output. Given a set of reference subtopics $\mathcal{G}_r$ and system-covered subtopics $\mathcal{G}_s$:

$$\text{S-Recall} = \frac{|\mathcal{G}_s \cap \mathcal{G}_r|}{|\mathcal{G}_r|} \tag{12}$$

The metric ranges from 0 (no subtopic covered) to 1 (all covered), and is agnostic to how subtopics are defined.

α-nDCG. Proposed by Clarke et al. (2008), α-nDCG extends normalised discounted cumulative gain to reward both relevance and novelty. For a ranked list of n items, the gain from covering subtopic g diminishes with each prior occurrence:

$$\alpha\text{-DCG} = \sum_{i=1}^{n} \frac{\sum_{g \in \mathcal{G}_i} J(g) \cdot (1-\alpha)^{c_{i-1}(g)}}{\log_2(i+1)} \tag{13}$$

where $J(g)$ is 1 if g is a relevant subtopic and 0 otherwise, $c_{i-1}(g)$ counts prior occurrences of g, and $\alpha \in [0, 1]$ controls the redundancy penalty ($\alpha = 0$: no penalty; $\alpha = 1$: only first mention counts). The score is normalised by an ideal ranking in which each relevant subtopic appears exactly once in the top positions. We use $\alpha = 0.5$.

Statistical Parity. A group fairness criterion from Dwork et al. (2012), Statistical Parity requires that outcomes be distributed proportionally across protected groups. Given representation rates r_g for each group g, the parity ratio is:

$$\text{SP} = \frac{\min_g \; r_g}{\max_g \; r_g} \tag{14}$$

A value of 1 indicates perfect parity. The widely-used *80% rule* (or four-fifths rule), originating from U.S. employment discrimination guidelines, considers a process fair if SP ≥ 0.8.