Scientific Reports

https://doi.org/10.1038/s41598-026-68010-x

Article in Press

A large-language-model framework for automated humanitarian situation reporting

Received: 3 December 2025

Accepted: 19 August 2026

Published online: 28 August 2026

Cite this article as: Decostanzi I., Mejova Y. & Kalimeri K. A large-language-model framework for automated humanitarian situation reporting. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-68010-x

Ivan Decostanzi, Yelena Mejova & Kyriaki Kalimeri

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

A Large-Language-Model Framework for Automated Humanitarian Situation Reporting

Ivan Decostanzi[1], **Yelena Mejova**[1], **and Kyriaki Kalimeri**[1,2,*]

[1]ISI Foundation, Turin, Italy
[2]UNICEF, NYC, USA
[*]kyriaki.kalimeri@isi.it

ABSTRACT

Timely and accurate situational reports are essential for humanitarian decision-making, yet current workflows remain largely manual, resource intensive, and inconsistent. We present a fully automated framework that uses large language models (LLMs) to transform heterogeneous humanitarian documents into structured and evidence-grounded reports. The system integrates semantic text clustering, automatic question generation, retrieval augmented answer extraction with citations, multi-level summarization, and executive summary generation, supported by internal evaluation metrics that emulate expert reasoning. We evaluated the framework across 13 humanitarian events, including natural disasters and conflicts, using more than 1,100 documents from verified sources such as ReliefWeb. The generated questions achieved 84.7 percent relevance, 84.0 percent importance, and 76.4 percent urgency. The extracted answers reached 86.3 percent relevance, with citation precision and recall both exceeding 76 percent. Agreement between human and LLM based evaluations surpassed an F1 score of 0.80. Comparative analysis shows that the proposed framework produces reports that are more structured, interpretable, and actionable than existing baselines.

By combining LLM reasoning with transparent citation linking and multi-level evaluation, this study shows that generative AI can substantially support the generation of humanitarian situation reports that domain experts judged more useful, better structured, and more transparently sourced than current approaches—marking a concrete step toward operational use, even as important challenges remain on the path to fully reliable autonomous deployment.

ARTICLE IN PRESS

Introduction

The humanitarian sector operates in complex, high-pressure environments characterized by rapidly changing circumstances, data scarcity, and fragmented information flows. In crisis situations such as natural disasters, conflicts, or health emergencies, organizations must collect, verify, and synthesize information from multiple heterogeneous sources, including government bulletins, NGO publications, news outlets, and internal reports, often under severe time constraints. One of the key outputs of this process is the *situational report*, a critical document that provides a concise snapshot of an ongoing humanitarian crisis, outlining its scope, impact, immediate needs, and ongoing relief efforts[1]. These reports are vital tools for coordination and decision-making among governments, NGOs, and international organizations. Yet, despite their importance, there is limited understanding of how artificial intelligence can replicate the analytical reasoning required to synthesize fragmented crisis information.

Producing high-quality situation reports remains a significant challenge. The process is typically manual, time-consuming, and resource-intensive, requiring substantial human effort to filter, summarize, and contextualize information derived from diverse and unstructured data sources[2]. Even with such investment, recent findings indicate that many reports, particularly within the UN system, are seldom read by their intended audience[3]. This intended audience is, moreover, far from uniform: situation reports are read by many different actors, who differ in role, background, and objective, and even domain experts vary in what they regard as valuable, relevant, or timely. Judging a report's quality in absolute terms is therefore difficult, since quality is multi-faceted rather than a single, fixed property — a view reflected in sector-wide quality and accountability standards developed beyond the UN system, such as the Core Humanitarian Standard[5]. These challenges underscore a fundamental research question: *can generative language models not only accelerate information synthesis but also preserve the factual reliability and analytical depth required for humanitarian operations?* Beyond accelerating synthesis, addressing this challenge is also a matter of how information is accessed: an effective system must let analysts reach the evidence relevant to a given decision and trace it back to its source, rather than read every document in full.

Recent advancements in Artificial Intelligence (AI), and particularly in Generative AI, show great promise in addressing these challenges[4]. AI systems can process and analyze large volumes of textual and numerical data, enabling faster and more accurate extraction of key insights from multiple sources. The emergence of large language models (LLMs) has transformed

ARTICLE IN PRESS

natural language processing (NLP), substantially improving performance in text classification, clustering, summarization, and question answering[6,7]. Because of their ability to model complex relationships in unstructured text, LLMs are increasingly explored for automating analytical tasks in humanitarian contexts, such as crisis-event classification and early-warning signal detection[7]. Nonetheless, it remains unclear whether these models can generate *verifiable, domain-grounded analytical outputs* that meet operational standards in humanitarian work.

Simply prompting LLMs to generate situation reports is often insufficient: models lack access to real-time or specialized data, may produce unverifiable or inconsistent statements, and offer limited control over output structure. Although several approaches exist for automating situation report generation, most lack a clear, easy-to-read structure that enables analysts to quickly grasp essential information. Many systems produce generic summaries[8,11] without mechanisms to assess or improve their quality. SmartBook[12] represents a notable step toward structured report generation, but it remains limited to news-like sources, omits features crucial for situational reporting (such as multi-level summaries and topic-level organization), and does not incorporate internal quality-evaluation mechanisms. These limitations are especially consequential in humanitarian contexts, where accurate and transparently sourced information is essential for decision-making affecting vulnerable populations.

Beyond the research literature, the humanitarian practice space has itself begun to adopt AI-driven tools for analysis and reporting. The Data Entry and Exploration Platform (DEEP), an open-source collaborative environment maintained by a consortium of humanitarian organizations, applies NLP models to assist the extraction and classification of secondary humanitarian data[9]. More recently, generative-AI assistants such as GANNET, developed by Data Friendly Space, have used retrieval-augmented generation with source citations to produce situational analyses and to monitor ongoing crises[10]. Such operational and emerging systems reflect a clear practitioner demand for automated situational awareness. However, the public literature provides limited information regarding their internal implementation details and evaluation procedures, making systematic comparison difficult. Our contribution is complementary to this practice landscape: rather than another closed tool, we present an openly specified and systematically evaluated pipeline whose individual stages—question generation, retrieval-augmented answering, citation post-processing, and summarization—are assessed both automatically and by domain experts.

Beyond the humanitarian sector, related AI-driven frameworks are being developed in areas such as climate monitoring and public health, where automated synthesis of complex, multi-source data supports rapid situational awareness[7,13,14]. However, few of these systems systematically evaluate generated content against human expert assessments or integrate explicit quality-control mechanisms. Building upon these advances, we propose and evaluate a modular framework for automated humanitarian report generation that emulates human analytical workflows through semantic clustering, retrieval-augmented question answering, and structured summarization. Our system extends prior designs such as SmartBook by incorporating *automatic internal evaluation metrics*, including domain-specific quality assessment[15–17], and by introducing visualization modules that contextualize results in terms of the Sustainable Development Goals (SDGs)[18]. The SDGs are, however, only one of several ways to organize humanitarian information, which in practice is often categorized by sectors or clusters such as water, sanitation and hygiene (WASH), health, or protection[30]. The pipeline is designed to mirror the natural workflow of human analysts: segmenting and clustering content into subtopics, generating targeted questions, and producing evidence-backed answers with citations to source material[12,20,21]. Because what counts as a high-quality report is multi-faceted, we evaluate the system along several complementary dimensions and draw on a range of evaluators, while recognizing that many further aspects of quality could be assessed. To ensure external validity and operational relevance, we collaborated with five experts from the United Nations Children's Fund (UNICEF) and Data Friendly Space (DFS), organizations that actively develop AI-powered tools for humanitarian analysis. These experts provided feedback throughout system design and participated in evaluating both intermediate components and final outputs. Comparative assessments against alternative systems show that humanitarian professionals consistently judge our framework to produce higher-quality, more actionable, and easier-to-navigate reports. Furthermore, we investigate the use of LLMs as automated judges in humanitarian NLP tasks, comparing their evaluations with expert assessments. Results show strong alignment across multiple dimensions, underscoring the potential of LLMs as reliable evaluators in this domain[16,24].

This work makes four primary contributions that advance the state of automated humanitarian information analysis: (i) we develop a generalizable pipeline applicable to any humanitarian crisis, natural or man-made, capable of handling diverse document types; (ii) we introduce a fully automatic, end-to-end reporting system that operates without human intervention, from document ingestion to structured report generation; (iii) we conduct extensive evaluation of each pipeline component with humanitarian experts and demonstrate strong correspondence between human and LLM-based quality assessments; and (iv) we publicly release all code and prompts used to construct our framework, ensuring transparency, reproducibility, and community-driven improvement.

ARTICLE IN PRESS

Methods

The goal of our framework is to transform raw humanitarian documents into structured, event-specific situational reports through a combination of semantic clustering, question answering, summarization, and visualization components. The methods described in this section detail the dataset used in the study, the design and implementation of the pipeline, and the evaluation procedures adopted to assess the quality and reliability of the generated outputs.

Dataset

The development of the pipeline and its evaluation are conducted on a dataset constructed from a large corpus of humanitarian documents provided by our collaborators at Data Friendly Space (DFS) under an established Memorandum of Understanding. The corpus compiled by DFS contains content from 142 distinct online sources, with approximately 40% originating from *ReliefWeb*, the leading humanitarian information service of the United Nations Office for the Coordination of Humanitarian Affairs (OCHA). In addition to ReliefWeb, the dataset includes material from other humanitarian and disaster-related platforms such as *The New Humanitarian* and *PreventionWeb*, as well as a broad range of global and regional news outlets, including *The Guardian*, *AllAfrica*, *Al Jazeera*, and *BBC News*.

As case studies for the situational reports, we define an *event* as an occurrence anchored to a specific geographic area and a time window (one week). We adopt a one-week window as a deliberate trade-off, consistent with the role of a situational report as a snapshot of an evolving crisis at a given moment: it is long enough for a meaningful volume of reporting to have accumulated, yet short enough to correspond to a concrete point in time at which operational decisions must be made. A comparable short, fixed window is used in related literature[11,12]. The window is a configurable parameter rather than a fixed requirement, and can be widened or narrowed to match the tempo of a particular situation. For protracted or ongoing crises—such as the Ukraine war, the Israel–Hamas war, or the Sudan conflict—an event corresponds to a one-week snapshot of the situation at a specific time, and the same crisis may be represented by several snapshots taken at different periods (as for the Sudan conflict and the Israel–Hamas war in Table 1a). In each case the window was chosen to fall on a period of heightened activity and dense reporting, ensuring sufficient textual evidence for analysis. Event-specific subsets of documents are created by filtering the corpus along these two dimensions. For each document, the referenced country is automatically inferred using a domain-specific geolocation tool[64] and subsequently validated through manual inspection. All documents whose publication date and associated location fall within the spatio-temporal scope of a given event are included in its corresponding subset. The complete list of events, along with the corresponding number of documents for each, is presented in Table 1a. The number of documents per event ranges from 47 to 143, resulting in a final dataset of 1,117 documents.

Pipeline

Starting from the collected documents, our pipeline proceeds in six main stages:

- **Semantic text clustering:** the documents are first segmented into smaller text units and grouped into clusters using HDBSCAN[25]. This allows semantically related content to be analyzed together.
- **Question generation:** to facilitate exploration of the clustered content, we automatically generate questions using LLMs[12,21].
- **Answers extraction:** information is automatically extracted from the clustered text segments through a Retrieval-Augmented Generation (RAG) pipeline[26].
- **Summary generation:** we generate (i) cluster-level summaries that synthesize all answers associated with each cluster, and (ii) SDG summaries that aggregate answers across clusters for all questions mapped to the same Sustainable Development Goal (SDG)[18].
- **Executive summary generation:** in parallel with the QA process, a concise summary of the situation is produced to provide readers with a clear and accessible overview at the beginning of the report.
- **Report Visualization:** Finally, we allow for four modes of visualisation, to facilitate the end users in operationalising the output of the framework, tailoring it to their specific needs.

An overview of this pipeline is presented in Figure 1. The following sections describe each stage in detail.

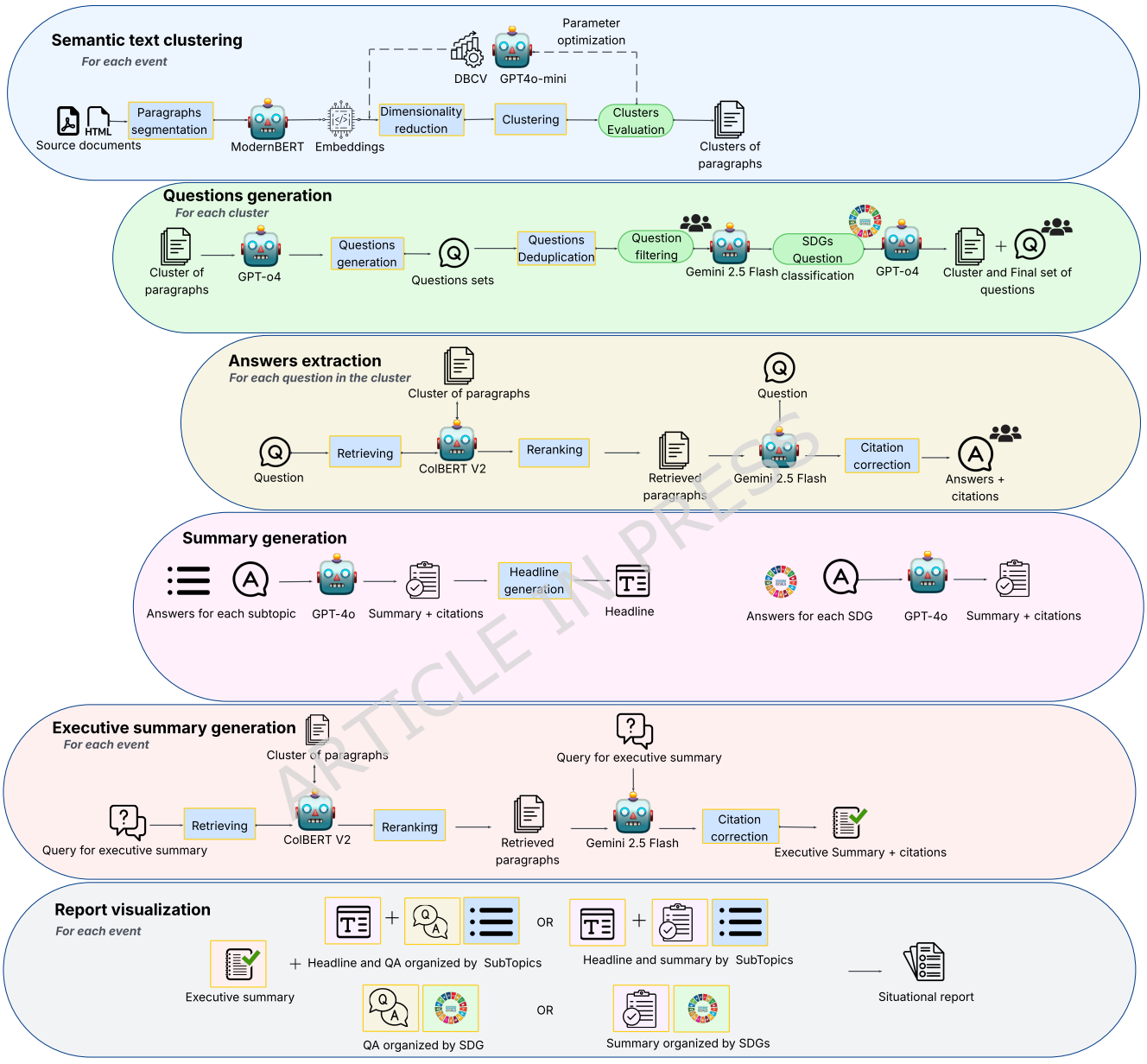


Figure 1. Overview of the proposed pipeline for transforming humanitarian documents into structured situational reports. The process comprises five main phases: semantic text clustering, automatic question generation, automatic answer extraction, creation of the summaries for each topic and for each SDG, executive summary generation,followed by a visualization module enabling four complementary modes of report exploration. Green boxes indicate components evaluated through automated metrics, while icons depicting human figures denote stages assessed by domain experts.

ARTICLE IN PRESS

Semantic Text Clustering

Initially, we focus on organizing the collected documents into meaningful thematic groups via semantic clustering.

Text Preprocessing. Unlike news articles, general humanitarian documents present unique challenges: they are often lengthy, hierarchically structured, and cover multiple interrelated topics, which limits the effectiveness of conventional document-level clustering approaches such as the one adopted in previous works[12]. To address this, we first segment documents into smaller sections prior to clustering. This process introduces the well-known challenges of short-text clustering, including data sparsity, high dimensionality, and limited contextual information[27,28]. After splitting the documents into paragraphs composed of four sentences each to ensure consistent and manageable text segments, we perform a series of standard steps[27] to clean and normalize the text: we replace encoded characters like smart quotes, dashes, and non-breaking spaces with standard ASCII equivalents and remove unwanted elements such as bullet points, tabs, and URLs while standardizing punctuation.

Embedding Generation. We embed the normalized paragraphs into dense vector representations employing the 768-dimensional embeddings produced by *ModernBERT*[29], a more efficient and higher-performing transformer encoder than the standard BERT model. In particular, we use the `nomic-ai/modernbert-embed-base` model[31] implemented in HuggingFace's Transformers library[32], leveraging its speed and strong semantic representation capabilities. Although LLMs can produce high-quality embeddings[6], ModernBERT offers an efficient and robust alternative, providing stable semantic representations at a fraction of the computational cost and it has a good placement in the MTEB Benchmark[33].

Dimensionality Reduction. High-dimensional embeddings often contain redundancies or noise that can hinder the clustering process, particularly when working with short texts[28]. To mitigate this issue[34], we employ UMAP (Uniform Manifold Approximation and Projection)[35] to decrease the number of dimensions to $k = 10$ (with k chosen after manual exploration). UMAP is often preferred over methods like PCA and t-SNE due to its balance of speed, scalability, and ability to preserve both global and local data structure[36].

Clustering. For the clustering phase, we employ the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm[25]. HDBSCAN offers several advantages essential for our task. [37] First, it automatically determines the number of clusters, which is particularly suitable when the number of distinct underlying sub-topics is unknown a priori. HDBSCAN was preferred to partition-based approaches such as k-means because the number of underlying humanitarian sub-topics is not known in advance and varies considerably across events. Further, it effectively identifies clusters of varying densities and constructs a cluster hierarchy, providing richer insights than a single flat partition. Its inherent ability to handle noisy data and robustness with minimal parameter tuning make HDBSCAN a well-suited choice for grouping paragraph embeddings.

Hyperparameter Search. The quality and interpretability of clustering results critically depend on hyperparameters in both the dimensionality reduction and clustering stages. To navigate this space, we employ a random search strategy, exploring combinations of UMAP's `n_neighbors` and `min_dist`, along with HDBSCAN's `min_cluster_size` and `min_samples`. Parameter selection is guided by automatic evaluation metrics. Specifically, we combine two complementary measures with a shared scoring range $[-1, 1]$. First, we adopt the Density-Based Clustering Validation (DBCV) index[38], specifically designed to assess the quality of density-based cluster structures[37]. Second, we introduce an LLM-based score, where *GPT-4o-mini* evaluates each cluster according to (i) *semantic coherence* and (ii) *homogeneity*[16,17], returning an overall score in the same range as DBCV. We choose to employ LLMs as judges since a growing body of evidence shows that they perform reliably in evaluation tasks[24,39]. The prompt used for this LLM-based evaluation is reported in the Supplementary Material S.0.1.1. The final hyperparameter configuration is selected by maximizing the mean of the DBCV and LLM-based scores, ensuring clusters are both structurally valid and semantically coherent[40].

Question Generation

Automatic Question Generation (AQG)[20] has emerged as a valuable, fast, and cost-effective approach for automatically identifying and extracting the most pertinent information from large sets of documents, regardless of the topic[21]. Instead of relying on manually defined or static questions, our system dynamically identifies those that are most relevant for situational analysis as the context evolves. We use this method to automatically detect crucial events, trends, and significant developments within each text cluster we previously generate. The aim is to present this key information in the form of questions, with their answers providing a concise summary. However, AQG alone does not guarantee that these questions are suitable for a situational report, particularly regarding their relevance to the topic and their appropriateness for the humanitarian sector[21]. To address this, we've developed a pipeline that filters and refines the generated questions, ensuring they meet the specific needs of a humanitarian report.

Prompt Design. To inform the design of prompts for generating questions suitable for humanitarian situational reports, we first conduct consultations with humanitarian experts. Experts emphasized that suitable questions should be non-political, focus on urgent, time-sensitive information rather than long-term implications, and be precise and concrete. Leveraging these insights, we develop three distinct prompts to investigate the impact of prompt design on question generation quality:

ARTICLE IN PRESS

- *Simple Prompt*: A simple prompt with fundamental instructions and rules designed to filter out poorly formulated questions.
- *Chain of Thoughts (CoT)*: Builds on the Simple Prompt by integrating a Chain-of-Thought (CoT) process[41] to guide the LLM towards more relevant questions through step-by-step reasoning.
- *CoT + Few Shot*: Extends Chain of Thoughts (CoT) by including examples to facilitate in-context learning, providing demonstrations of the desired question format and content.[42]

Based on our manual evaluation, all prompts performed similarly, but the Simple Prompt was selected because it yielded slightly better results and the added complexity of the other prompts offered no clear benefit for this straightforward task. The prompt can be found in the Supplementary Material S.0.0.1

Question Generation. Questions are generated by prompting *GPT-4o* [43] (as of February 2025), leveraging the established capability of LLMs to produce relevant questions on specific topics[21]. The clustered paragraphs are provided in the prompt to ensure contextual grounding. To promote diversity[44], we employ nucleus sampling[45] and prompt the model three times, producing three possibly redundant sets of questions. To remove redundancy, we use the RoBERTa-large model[46] (`cross-encoder/quora-roberta-base`), pre-trained on the Quora Duplicate Question Pairs dataset, which assigns a score between 0 and 1 indicating the likelihood of two questions being duplicates. Redundant questions are filtered out at the generation stage, and a final sample of six questions is taken.

Question Filtering. Even with explicit instructions, some generated questions may be unrelated to the main event due to hallucinations or irrelevant references (e.g., mentions of neighboring regions). To address this, we develop an automated filtering system using *Gemini 2.5 Flash*[47] as an LLM judge[48], guided by four binary filtering criteria based on expert feedback: *a) Not Specific to country*: References to locations unrelated to the primary focus. *b)Too Political*: Focus on political causes, opinions, or blame instead of humanitarian impact. *c)Long-term/Historical*: Emphasis on past events or future scenarios lacking current humanitarian relevance. *d)Vague/Irrelevant*: Overly broad, abstract, or excessively specific questions. A question passes the filter only if it receives a 1 for all four criteria. To reduce the so-called 'Self-Prevalence Bias' that can occur when an LLM judges its own output, which is a concern for models such as GPT-4o[49], we use a different model (*Gemini 2.5 Flash*) to evaluate and filter the quality of the answers. We assess the reliability of our judgments against human judgments, and the detailed results can be found in the Results section. The optimized prompt for the evaluation can be found the Supplementary Material. (see S.01.2).

Question Classification by SDGs. Beyond filtering, each retained question is categorized according to the United Nations Sustainable Development Goals (SDGs)[18,19]. We provide the model (again *GPT-4o*) with concise descriptions of all 17 SDGs (see S.0.14) and ask it to assign each question to one or more relevant goals. The prompt used for this classification is reported in Supplementary Material S.0.13 . We adopt the SDGs as one possible framework for classification, as they offer a recognizable cross-organizational reference point in humanitarian contexts[18]. Moreover, they offer a structured way to group questions that point in the same *direction*,for instance, those related to health, education, or climate resilience,thereby enhancing thematic coherence and interpretability in the reports.

Answer Extraction

Standard Large Language Models are prone to generating factual inaccuracies or "hallucinations", particularly when dealing with information absent from their training data[26,50]. Retrieval-Augmented Generation (RAG) addresses this by dynamically incorporating external knowledge[51]. The core RAG process involves retrieving relevant documents or passages from an external knowledge base in response to a query, and then providing this retrieved information as context to the LLM during the answer generation phase. This mechanism offers crucial advantages for tasks requiring accurate answer extraction. By grounding the LLM's generation process in specific, retrieved evidence, RAG significantly enhances the factual accuracy and relevance of the outputs, reducing the likelihood of fabricated information[13,51]. Additionally, RAG exhibits flexibility, proving effective not just for general knowledge-intensive tasks like Open-domain Question Answering (OpenQA), but also for handling diverse question types requiring specific contextual information[50]. To further enhance trustworthiness and transparency, RAG systems often incorporate citations or attribution[52], explicitly linking generated statements back to the supporting evidence within the retrieved source documents, which is crucial for enabling factuality verification and boosting model credibility[53]. We selected a retrieval-augmented approach because humanitarian reports require answers grounded in current event-specific evidence and traceable to source documents, properties that cannot be guaranteed by relying solely on model parameters.

Paragraph Retrieval. Given these advantages, we employ a RAG pipeline using ColBERTv2[54] for indexing and retrieving paragraphs from our document database. ColBERTv2 employs a late interaction mechanism that enables efficient token-level matching between queries and documents and despite being introduced in 2020, ColBERT remains a widely adopted and efficient dense retrieval model[51].

ARTICLE IN PRESS

Reranking and Generation. To enhance retrieval quality, we implemented RAG-Fusion[55], which generates multiple query variations and applies reciprocal rank fusion for robust reranking[50]. The reranked documents, together with the original query and tailored instructions, are passed to the *Gemini 2.5 Flash* model to synthesize answers strictly based on the retrieved context, with instructions to include citations referencing the source documents for each piece of the answer. We selected this model as it is freely accessible and yielded superior answer quality in preliminary experiments on our humanitarian corpus compared with the alternative models we evaluated, including *GPT-4o*. This model is selected due to its strong performance across various tasks and its favourable placement in the chatbot arena[56]. The prompt used is available in the Supplementary Material S0.1.5
Post-processing Citations. To address instances of incorrect citations, we developed a post-processing correction method. While many studies have demonstrated the effectiveness of such approaches, they often rely on fine-tuned models, which is not feasible in our setting due to limited resources and time constraints[52]. Instead, we drew inspiration from a recent work proposing several lightweight and efficient post-processing techniques[57]. Among these, a 'keyword and semantic context-based matching' approach proved most effective. This method corrects citations by re-evaluating the relevance of retrieved documents to each claim in the LLM's response. The answer is segmented into discrete claims (x_i), each identified by their associated citations using regular expressions to detect citations at the end of factual statements. For each claim (x_i), the goal is to identify the most relevant retrieved document ($\hat{x}_j$) by employing a similarity metric (s_{ij}) that must exceed a predetermined threshold. The similarity is computed as: $s_{ij} = \lambda \cdot J(x_i, \hat{x}_j) + (1-\lambda) \cdot \text{cosine_similarity}(q, \hat{x}_j)$, where $J(x_i, \hat{x}_j)$ denotes the Jaccard similarity between the claim and the retrieved document, measuring keyword overlap, and $\text{cosine_similarity}(q, \hat{x}_j)$ is the cosine similarity between vector embeddings of the original query q and the document $\hat{x}_j$, capturing semantic relevance. We empirically set $\lambda = 0.8$, emphasizing direct keyword matching while incorporating broader semantic context. The pipeline is implemented using LangChain to facilitate seamless model integration and customization.

Summary Generation

To support multi-level analysis of the extracted information, the system produces two different types of summaries. First, for each cluster, we generate a consolidated summary that synthesizes all answers associated with that sub-topic. Second, for each SDG, we generate thematic summaries that aggregate answers across clusters for all questions mapped to that SDG. During summarization, the large language model *GPT4o* is explicitly instructed to preserve all citations associated with each claim, and the full prompt is provided in the Supplementary Material. The summary is inputted to *GPT4o* which provides with a headline that describe the topic.

Executive Summary Generation

The final step of our pipeline consists of generating an executive summary that provides a concise, high-level overview of the main findings. This summary is constructed exclusively from the paragraphs used to generate the Q&A answers, ensuring that it remains grounded in the same evidence base. To maintain transparency and credibility, citations are preserved and linked back to the original source documents. For this purpose, we implement a retrieval-augmented generation (RAG) pipeline, similar to the one described in the previous section, but applied only to the subset of paragraphs used in the answers rather than to the full corpus. The specific prompt used for the executive summary generation is provided in the Supplementary Material S0.1.6

Report Visualization

To enhance readability and usability, the final report is presented through four complementary visualization formats. Specifically, we organize and display the question–answer outputs in the following ways: (1) grouping questions and answers by sub-topic; (2) grouping questions and answers by SDGs; (3) presenting cluster-level summaries; and (4) presenting SDG-level summaries. These visualizations allow users to explore the content according to their analytical needs,either by examining detailed, fine-grained responses or by consulting higher-level thematic summaries.

Established humanitarian situation reports typically follow fixed, sector-organized templates defined by operational organizations such as OCHA and CARE[22,23]. Our report deliberately departs from these fixed formats: rather than reproducing a standard template, it adopts an exploratory, evidence-linked structure — question–answer pairs, multi-level summaries, and multiple organizational modes with traceable citations — in the spirit of prior AI-assisted situation-reporting work[12], designed to help analysts navigate and verify the underlying evidence. This organization was refined through feedback from senior humanitarian experts at UNICEF and Data Friendly Space (DFS) to support efficient expert navigation.

Finally, all citations are reindexed to ensure unique numbering and consistency across the entire report, merging duplicates and maintaining coherence across all visualization formats and the executive summary. An example of the four report formats is shown in Figure 2. The reports for the various events can be visualized at this link.

Evaluation

We thoroughly assess every step of the pipeline, focusing on the quality of the generated questions, the quality of the answers and their associated citations, and the overall quality of the final report. Because the quality of a report is multi-faceted and depends on the reader's role and objectives, as outlined in the Introduction, we deliberately assess it along several complementary

ARTICLE IN PRESS

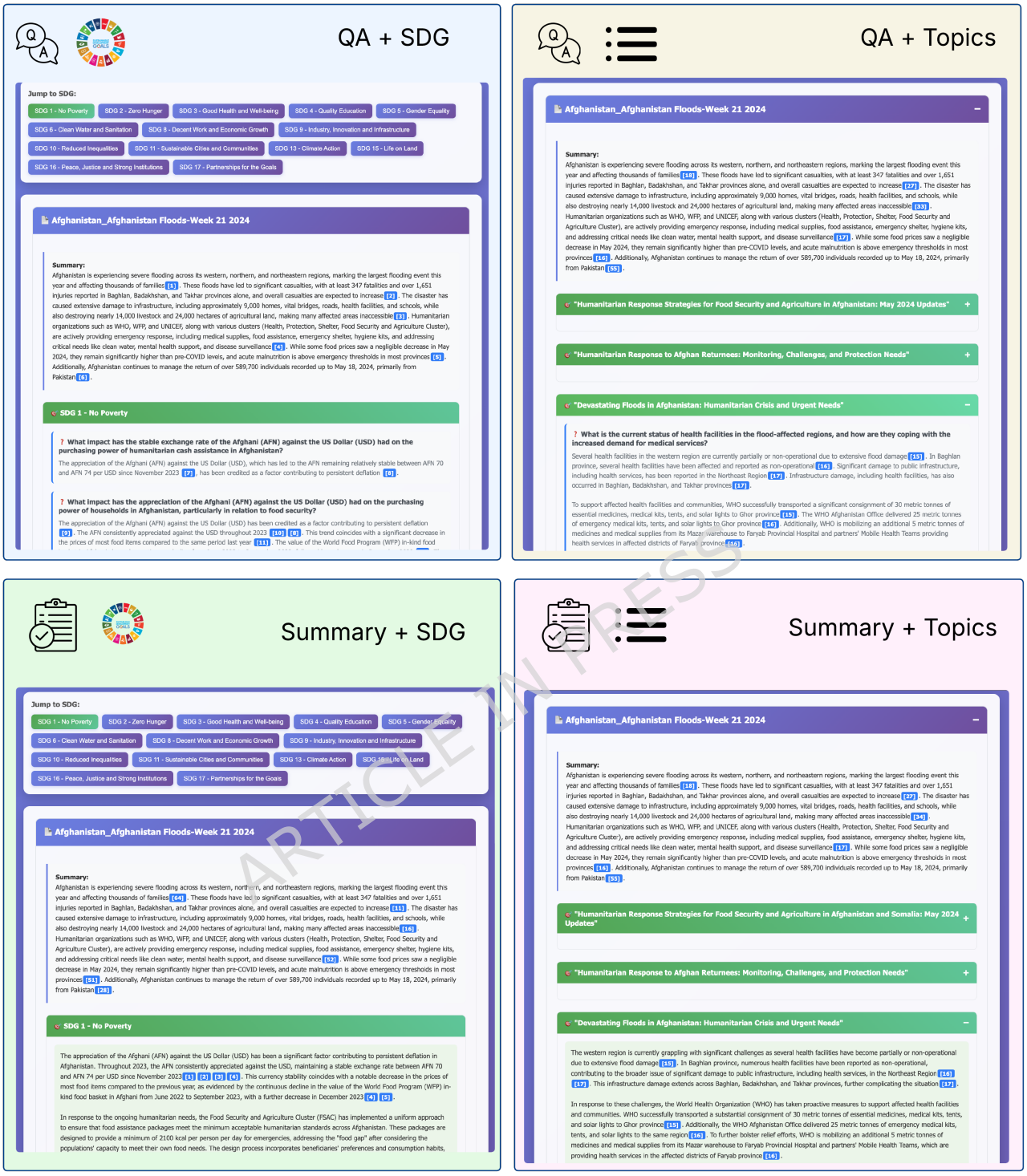


Figure 2. This figure illustrates that our pipeline produces four different versions of the situational report, allowing the user to choose which one to view. The reports can be visualized either organized by Sustainable Development Goals (SDGs) or by automatically discovered sub-topics. Each version can be displayed in the form of a Q&A or as summarized text. All four versions include an executive summary at the beginning and provide citations for the specific claims made. Each citation includes the original paragraph, the document title, and a link to the source document.

ARTICLE IN PRESS

dimensions rather than through a single overall score. To ensure a comprehensive assessment, the evaluations were carried out by three distinct groups, matched to the nature of each task. The three authors performed the technical selection and validation steps — comparing the candidate question-generation prompts and validating the automatic filter against human judgment. A panel of humanitarian experts performed the two tasks that require domain judgment — the final assessment of question quality and the overall comparison of the reports. A separate group of trained annotators carried out the answer- and citation-verification task, which relies on careful checking against the source documents rather than on operational humanitarian expertise. This multi-perspective approach allows us to capture both domain-specific insights and unbiased external judgments. The study received ethical approval from the ISI Foundation Ethical Review Board (IRB) and was conducted in compliance with institutional policies and applicable data protection regulations. No sensitive personal data were collected or processed, and expert evaluators participated voluntarily after being informed about the study procedures. The full set of results are available upon request, while additional details about the expert participants and the evaluation guidelines can be found in the Supplementary Information. No sensitive personal data were collected beyond professional feedback.

Evaluation of Question Generation and Filtering. We evaluate the automatic question generation process on three aspects: (i) prompt optimization, (ii) the effectiveness of the filtering mechanism, and (iii) the quality of the final questions after filtering. To find the best prompt for question generation, we manually examine the outputs of the three prompt designs (Prompt 1–3) to assess their suitability for humanitarian situational reports, using the criteria of *Relevance* and *Importance*. Here, *Relevance* (standard measure in Information Retrieval[58]) reflects whether the question pertains directly to the specific humanitarian event, and *Importance* captures whether the question is valuable for inclusion in a humanitarian situation report. We are especially interested in the latter measure of the kind of information that is most useful to humanitarian efforts, beyond simple relevance. To ensure an unbiased evaluation, we randomly select 33 questions from the topics generated by each prompt (totalling 100 questions overall) and assess their quality without knowing which prompt had produced them. The evaluation is conducted independently by the three authors, with partial overlap in their annotations to allow for consistency checking and inter-annotator agreement assessment.

For the filtering of the generated questions, we assess GPT-4o's performance on binary classification tasks concerning the four metrics outlined in the Question Filtering subsection against human annotations that are performed by the authors. We summarize its performance by computing precision, recall, and F1-score. This evaluation follows established practices for assessing LLMs as judges[16,17]. In this phase, 200 questions were randomly sampled from the set of generated events and annotated according to the four metrics, again involving the three authors with overlapping assignments to ensure consistency and reliability.

The quality of the final set of generated questions, obtained after filtering and using the optimal prompt strategy, is assessed by three humanitarian experts according to three binary criteria: *Relevance* and *Importance*, as defined above, as well as *Urgency*, which captures whether the question addresses short-term needs such as food, water, shelter, health, or protection. In total, 150 questions were evaluated in this phase, with some overlap across annotators to support consistency assessment. The *Urgency* dimension is introduced by the humanitarian experts during the exploratory feedback session to better capture immediate humanitarian priorities. Information about the annotators can be found in the Supplementary Material S4.

We compute inter-annotator reliability to measure the consistency of judgments both among human annotators and, where appropriate, between human annotators and the LLM output. Specifically, we calculate the Percentage of Agreement, Cohen's Kappa[59], and the Prevalence-Adjusted Bias-Adjusted Kappa (PABAK), which is less sensitive to the prevalence paradox - the phenomenon where Kappa can appear artificially low when the distribution of categories is highly imbalanced, even if agreement between annotators is high. [60,61]. The same inter-annotator reliability analysis is conducted for the evaluation of answers (see the next Section). The detailed annotation guidelines for each of the above tasks can be found in the Supplementary Material S0.2.

Evaluation of Generated Answers. To evaluate the quality of the generated answers for the questions that passed our filtering system, we conduct an extensive manual annotation effort involving 13 trained annotators — independent researchers, each holding a master's or doctoral degree — who followed a shared set of annotation guidelines. As this stage assesses answer relevance and whether each statement is supported by its cited source, it is primarily a verification task and does not require operational humanitarian expertise; consistency was instead ensured through the common guidelines and verified through inter-annotator agreement. The evaluation focus on two primary metrics: *Answer Relevance*[62] and *Citation Quality*[52,63]. We randomly select 10 question–answer pairs from each of the 13 humanitarian events, resulting in a total of 130 evaluated answers. These are distributed among the 13 annotators with partial overlap to allow for consistency checking and inter-annotator reliability assessment. Across this sample, annotators evaluate 983 individual claims and 1,146 supporting citations. To ensure the robustness of the evaluation, we compute inter-annotator agreement to verify the consistency of judgments among annotators.

We assess *Answer Relevance* by determining whether the generated answer provided information directly pertinent to the posed question. A binary label was used: relevant or not relevant[62]. *Citation Quality* is evaluated through *Citation Recall*

ARTICLE IN PRESS

and *Citation Precision*, following definitions from prior work[52,63]. Specifically, for each answer, composed of n statements $s_1, \ldots, s_n$, each statement s_i cites a set of passages $C_i = \{c_{i,1}, c_{i,2}, \ldots\}$.

- *Citation recall* evaluates whether the entire statement s_i is supported by its associated set of cited passages C_i.
- *Citation precision* measures whether each individual citation in C_i is relevant to the statement s_i.

To operationalize this for human annotators, a three-point scale is adopted for both citation recall and precision:

- *Fully supports*: The citation (for precision) or the set of citations (for recall) entirely substantiates the statement s_i.
- *Partially supports*: The citation or set of citations offers partial but incomplete support for s_i.
- *Does not support*: No support for s_i is provided.

When a single citation is associated with a single statement, citation recall and precision are equivalent.

Usability Evaluation. After generating the final report, we assess its quality through a structured evaluation involving a panel of 5 humanitarian experts. The evaluation consists of three stages. In the first stage, experts examine the individual components of our system, providing feedback on their quality and their usefulness. In the second stage, they compare outputs from three systems,our proposed model, Smartbook[12], and a state-of-the-art multi-document summarization system based on large-context prompting[11],without being informed which system produced each output. Finally, in the third stage, the experts rank the systems overall and provide open-ended qualitative comments regarding their perceived strengths and limitations. The survey is designed following a five-point Likert scale adapted from prior work[12] to capture quantitative assessments of quality, coherence, and informativeness. To ensure a fair comparison, we standardize the underlying model architecture across systems, employing Gemini 2.5 Flash[47] for both Smartbook and the reference summarization baseline. This approach ensures consistent technological conditions across all experiments. The detailed survey instructions and question items are provided in Supplementary material S0.2.4.

Results

In summary, for a given event defined by a specific country and time period, our pipeline segments and clusters the document texts, generates questions with corresponding answers and citations, produces summaries at both the cluster and SDG levels and produces an executive summary at the beginning of the report. We then evaluate the system's performance at each stage of this process.

Semantic Text Clustering

The first step of the pipeline involves the topical clustering of the text segments from the documents relevant to the event. Table 1b reveals significant variation in the number of clusters across events. While some, such as the Afghanistan Floods and Indonesia's events, show a smaller number of clusters (7 each), others, for instance, the May Israel-Hamas war, have a much higher count. This wide range, from 6 to 32 clusters, highlights the diverse structural complexity of the data for each humanitarian crisis and this is not correlated to the amount of documents available (Pearson coefficient = 0.1). Interestingly, the combined semantic coherence and homogeneity score produced by the LLM-as-a-judge remains consistently high, generally above 0.83, even with this diverse clustering, suggesting the LLM's robust clustering performance across varying data structures. There is a moderate correlation between the Density-Based Clustering Validation (DBCV) and the LLM score, the Pearson coefficient between them is 0.4, sign that there is a relation between the metrics we used.

Question Generation

Prompt Design. We compared three approaches: a simple prompt, a Chain-of-Thought (CoT) prompt, and a CoT prompt enriched with few-shot examples. We manually evaluated a sample of 100 questions,roughly one-third from each prompt,for their relevance to the humanitarian context and overall importance. Inter-annotator agreement was computed using 15 shared questions per annotator, and we observed perfect agreement across all evaluations. The results show that the simple prompt performs slightly better than the more complex alternatives. It achieves the highest relevance score (85.1%), compared to 81.2% for CoT and 79.3% for CoT with few-shot examples. A similar pattern holds for importance: the simple prompt reaches 96.5%, outperforming both CoT (93.3%) and CoT with few-shot examples (93.6%). Given this evidence, we adopt the simple prompt in our final pipeline, described in Supplementary material S.0.1. .

Question Filtering. To assess the reliability of our Gemini based filtering system, we evaluate a dataset consisting of 200 questions generated before filtering, consisting of two questions for each cluster and for each event. These questions are then coded for the four filtering metrics: Relevance,Political,Long term, and Generality, as defined in the Methods section. The

ARTICLE IN PRESS

	Event	Period	Type	N Docs
1	Floods and volcanic activity in Indonesia	May 06 – May 12	Natural Disaster	61
2	Afghanistan Floods	May 13 – May 19	Natural Disaster	67
3	Cyclone Remal in Bangladesh	May 13 – May 19	Natural Disaster	53
4	Hurricane Beryl in Jamaica*	Jul 08 – Jul 14	Natural Disaster	47
5	Landslide and Floods in India	Jul 22 – Jul 28	Natural Disaster	84
6	Flooding in Nigeria	Sep 9 - Sep 15	Natural Disaster	98
7	Ukraine conflict	May 27 – Jun 02	Conflict/Violence	61
8	Israel-Hamas war*	May 6 - May 12	Conflict/Violence	122
9	Sudan conflict*	Aug 19 – Aug 25	Conflict/Violence	100
10	UK riots	Jul 29 – Aug 04	Conflict/Violence	86
11	Sudan conflict	Sep 16 – Sep 22	Conflict/Violence	126
12	Gang violence and humanitarian crisis in Haiti	Sep 23 – Sep 29	Conflict/Violence	69
13	Israel-Hamas war	Sep 30 – Oct 06	Conflict/Violence	143

(a)

	Event	Number of Clusters	DBCV	LLM Score
1	Floods and volcanic activity in Indonesia	7	0.685	0.943
2	Afghanistan Floods	7	0.411	0.938
3	Cyclone Remal in Bangladesh	11	0.360	0.964
4	Hurricane Beryl in Jamaica*	9	0.415	0.984
5	Landslide and Floods in India	16	0.402	0.835
6	Flooding in Nigeria	6	0.299	0.849
7	Ukraine conflict	15	0.552	0.860
8	Israel-Hamas war*	32	0.292	0.837
9	Sudan conflict*	6	0.318	0.903
10	UK riots	7	0.251	0.864
11	Sudan conflict	7	0.549	0.830
12	Gang violence and humanitarian crisis in Haiti	19	0.459	0.870
13	Israel-Hamas war	7	0.683	0.996

(b)

ARTICLE IN PRESS

Table 1. (a) Events used to test our framework with the period in which they occurred and the number of documents we have available for each event. Events marked with an asterisk (*) denote the three cases—two conflicts and one natural disaster —selected for the expert evaluation of question quality and used as benchmarks for the computational-efficiency analysis. (b) Clustering performance results for each event showing the number of clusters identified, DBCV (Density-Based Clustering Validation) score, and LLM evaluation score for homogeneity and coherence.

inter-annotator agreement, detailed in Supplementary Table S1, demonstrates good overall reliability in our human annotation process. Specifically, the PABAK scores indicate substantial agreement for Relevance (0.867) and Political (0.822). Agreement for Long term (0.778) is also good, while Generality (0.644) shows moderate agreement. As anticipated, Cohen's Kappa values were lower across the categories, likely reflecting the impact of the prevalence paradox within our dataset, however, the robust PABAK scores confirm the consistency of our human judgments. The final labels are selected randomly where annotators disagreed.

Using this annotated dataset, we evaluate the performance of the Gemini 2.5 Flash filtering model by comparing its judgments against the aggregated human annotations. Table 2 presents the precision, recall, and F1 scores for this comparison. The system demonstrates notably high precision across all individual metrics, for instance, achieving 94.7% for Relevance and 93.3% for Political. This suggests that when the LLM identifies a question as meeting a specific criterion, it is highly likely to be correct. Recall scores are also strong, particularly for Generality (87.2%) and Political (85.9%), indicating the LLM's effectiveness in identifying most of the instances of these categories. Consequently, the F1 scores, which balance precision and recall, are high for Generality (89.2%) and Political (89.4%), and also reflected strong performance for Long term (86.5%)

ARTICLE IN PRESS

and Relevance (86.3%). For the overall decision of whether to *"Keep the question"*, the system achieves a balanced F1-score of 79.1%, underscoring good practical alignment with human consensus. To quantify statistical uncertainty, we computed 95% confidence intervals (CIs) for precision, recall, and F1-scores across the four filtering metrics using bootstrap resampling ($n = 1{,}000$). The mean F1-score for the overall decision was 79.1% (95% CI [73.4, 84.3]), while the individual metrics show similarly narrow intervals ranging from ± 3.9 percentage points for *Relevance* to ± 3.4 percentage points for *Generality*. These results confirm that performance differences across criteria are statistically meaningful rather than due to sampling variability.

Category	Agreement (%)	Precision (%)	Recall (%)	F1 (%)
Relevance	77.7 [71.8, 83.2]	94.7 [91.0, 98.0]	79.4 [73.3, 85.0]	86.3 [82.2, 89.9]
Political	82.1 [76.7, 87.1]	93.3 [89.1, 97.0]	85.9 [80.7, 90.3]	89.4 [85.9, 92.5]
Long term	78.2 [72.8, 83.7]	89.8 [85.0, 94.2]	83.5 [77.9, 88.8]	86.5 [82.5, 90.2]
Generality	81.2 [75.7, 86.2]	91.3 [86.8, 95.2]	87.2 [82.3, 91.8]	89.2 [85.6, 92.4]
Keep the question	74.2 [68.3, 80.2]	82.4 [75.9, 88.6]	76.2 [69.0, 83.3]	79.1 [73.4, 84.3]

Table 2. Agreement between human and LLM scores for the quality of the generated questions, expressed in percentages with 95% confidence intervals in brackets.

Question Generation. The quality of the final set of generated questions (after using the optimal prompt strategy and filtering) is assessed by three humanitarian experts based on Relevance, Importance, and Urgency. Considering limited time availability, we select the three topics from the Table 1a marked with * for this expert evaluation, as they cover two different conflicts and a natural disaster. The experts evaluate 150 questions, with an overlap of 15 questions between them. Inter-annotator agreement for this final quality assessment is presented in Supplementary Table S2. The PABAK scores indicate fair to moderate agreement across the three metrics: Relevance (0.595), Importance (0.563), and Urgency (0.510). This suggests that while there is a general consensus, the subjective nature of these qualities can lead to some divergence in expert opinions. As with the filtering evaluation, Cohen's Kappa scores are lower, highlighting the effect of score distribution on this metric. Considering the performance of the system, Table 3 summarizes the scores provided by each expert for each of the metrics and their mean scores. On average, questions were rated highly for Relevance (84.7%) and Importance (84.0%), and moderately for Urgency (76.4%). Expert 1 and Expert 3 provide consistently higher scores across Relevance and Importance compared to Expert 2, while scores for Urgency showed less variance among experts. These results point to the diversity of opinion possible even among experts.

Overall, the experts note that the evaluated questions contained a good mix of detailed and broader questions, with the detailed questions being particularly valuable for creating situational reports. They also observed that some topics produce higher-quality questions than others. Questions that are rated as less relevant typically addressed very specific topics mentioned in the source documents but are not essential for the situational report. For example, the question "What measures are being taken to ensure the safety and legal compliance of Cash for Work participants in Khyber Pakhtunkhwa, particularly concerning occupational safety and health standards?" refers to a program mentioned in the documents but is not directly relevant to a situational report focused on the floods in Pakistan. These observations highlight the diversity of opinion possible even among experts and point to areas where the system could be further refined.

Metric	Expert 1 (%)	Expert 2 (%)	Expert 3 (%)	Mean Score (%)
Relevance	92.7	72.7	88.7	84.7
Importance	94.5	67.3	90.3	84.0
Urgency	70.9	74.5	83.9	76.4

Table 3. Evaluation scores for the quality of the generated questions by the experts, expressed in percentages.

Answer Extraction and Citations

To evaluate the quality of the generated answers, we perform an extensive labelling effort resulting in 983 claims and 1146 citations annotated for 130 questions with the help of 13 independent annotators. The agreement on Answer Relevance is perfect among all of the annotators, whereas it was at 0.90 for Citation Precision and 0.81 for Citation Recall, as can be seen in Supplementary Table S3.

Table 4 summarizes the system's performance. We find that a large percentage 86.3% of the answers are deemed relevant. For citation quality, a substantial portion of citations are found to be relevant or partially relevant, both in terms of precision and recall. Specifically, for citation precision, 76.3% of citations are relevant and 5.93% are partially relevant, out of 1146 evaluated

ARTICLE IN PRESS

citations. For citation recall, 76.4% of claims are supported and 5.90% are partially supported, out of 983 sets of citations evaluated. This assessment highlights a strong overall performance of the model in both answer relevance and citation quality.

Metric	Count	Relevant (%)	Supported (%)	Partially Supported (%)
Answer Relevance	130	86.3	–	–
Citation Precision	1146	–	76.3	5.93
Citation Recall	983	–	76.4	5.90

Table 4. Quality assessment of generated answers and citations.

We also evaluate answer relevance using an LLM as judge[62], as we did in the previous section. To assess the reliability of this evaluation, we again employed bootstrap resampling ($n = 1{,}000$) to compute 95% confidence intervals. The agreement with human annotators was high, with a Percentage of Agreement of 70.1% (95% CI [67.1, 74.2]), Precision of 69.5% (95% CI [67.3, 72.9]), Recall of 97.1% (95% CI [95.3, 98.8]), and F1 Score of 81.0% (95% CI [79.6, 83.4]), confirming good overall alignment between human and LLM judgments, consistent with the results reported in Table 2 for the quality of the questions. Those classified as Not Relevant are typically answers where the model inferred a response even when the sources did not provide sufficient information, often using phrases such as *"The sources do not provide clear answers to that"*, followed by a speculative or tangential response. Another common issue arose when the answer closely mirrored the phrasing of the question without adding substantive information. For example: *"**Q:** What specific warnings have been issued to local aviation authorities regarding the volcanic activity of Mount Ibu?" "**A:** The country's volcanology agency issued a warning for aviation authorities managing local flights[41]."* In such cases, although the answer technically responds to the question, it does so with minimal value-added content, often lacking synthesis or analytical depth.

Turning to citations, we first consider citation recall, which evaluates whether all claims made in the answer are supported by the retrieved evidence. We find that 86.6% were fully or partially supported. Notably, in all cases where claims were *not* supported (13.37%), there was only one citation associated with the claim, it's the case in which citation recall is exactly the same as citation precision. Regarding citation precision, 86% of citations were judged as relevant or partially relevant, indicating a generally strong link between statements and their supporting references. However, the remaining 14% highlight important limitations. The primary sources of citation error are: (i) *Unnecessary citations*: Additional references included even when not needed. (ii)*Inferred citations*: The claim is generally supported by the cited sources, but not stated explicitly; it is derived by interpreting or combining information across multiple contexts. (iii) *Missing citations*: Statements are made without any explicit support in the retrieved documents. (iv) *Hallucinated content*: Entirely fabricated statements not grounded in any source material. Among these, hallucinations are particularly problematic, as they represent a breach of factual reliability. This issue has been widely discussed in recent literature on large language model behavior[53,65], and poses significant challenges for the deployment of such systems in sensitive or high-stakes domains. A manual analysis of citation errors shows that 5.1% of citations are classified incorrect, 4.5% were inferred, 4.3% as unnecessary, and 3.0% as hallucinations.

Overall Report Evaluation

We conclude by presenting the expert evaluation of the reports produced by our system, in comparison with two others. Across all stages of the final report evaluation, a total of five experts with direct experience in producing and interpreting humanitarian situational reports participated in the assessment, including four analysts from UNICEF and one technical specialist from Data Friendly Space. All were data scientists working on humanitarian analysis and reporting, with two to five years of sector experience and specializations spanning natural disasters, armed conflict, climate- and migration-related indicators, report authorship, and natural-language processing; per-expert details are given in Supplementary Material (Table S0.5). Experts 1–3 evaluated both the quality of the generated questions and the final reports, whereas Experts 4–5 took part only in this final evaluation; all five are distinct from the annotators involved in the answer-evaluation stage. In the first stage of evaluation, experts examined individual components of our system. They expressed strong appreciation for features unique to our approach, particularly the executive summary at the beginning of the reports and the ability to explore the report through multiple structures and visualizations. This flexibility, absent in SmartBook, was repeatedly highlighted as important for accommodating diverse analytical preferences. Experts also emphasized the importance of citation transparency in humanitarian reporting; this requirement is met by both our system and SmartBook, but is entirely missing from the summariser[52].

In the second stage, experts compared full reports from all three systems using a five-point Likert scale. The quantitative results are shown in Figure 3. Several clear patterns emerge from these evaluations. First, our system achieves strong performance in dimensions directly related to analytical quality. Experts rated it highly for its ability to identify and address strategic elements relevant to the situation, integrate information from multiple sources, and provide depth sufficient to support operational decision-making. In these categories, our system consistently exhibits the largest proportion of "Agree" and

ARTICLE IN PRESS

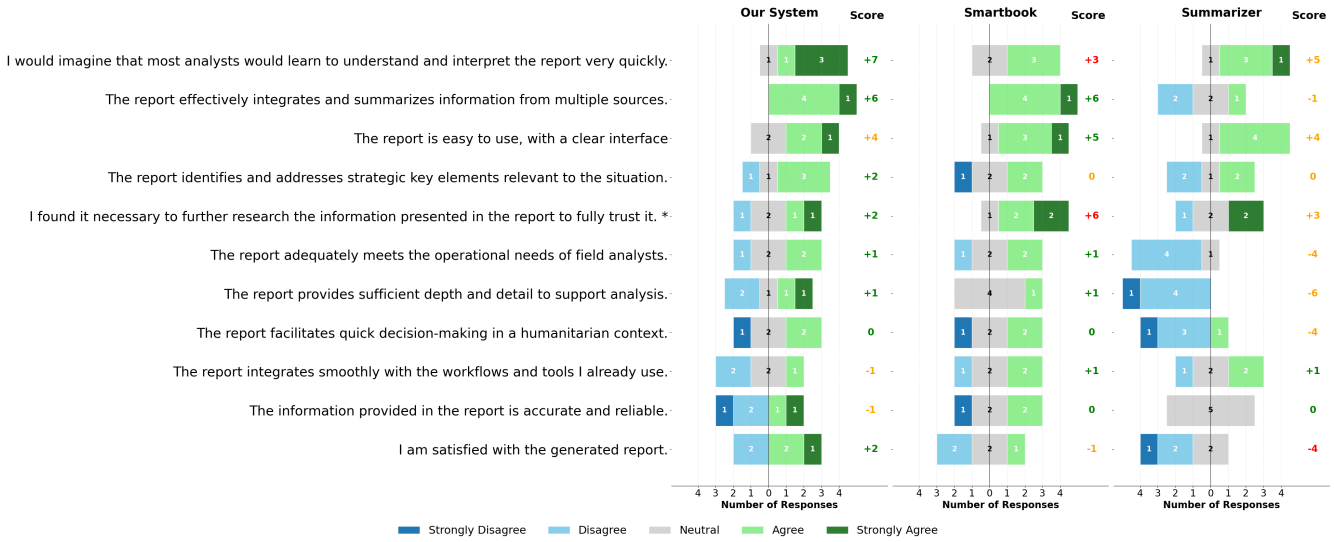


Figure 3. Expert evaluation of the three reporting systems: (i) our proposed framework, (ii) SmartBook, and (iii) a summarizer, across multiple dimensions of report quality. Evaluations were conducted by five humanitarian experts with direct operational experience in producing and interpreting situational reports. Responses are shown on a five-point Likert scale, ranging from "Strongly Disagree" (dark red) to "Strongly Agree" (dark green). For each criterion, coloured bars represent the number of expert ratings falling into each category. Higher concentrations of green indicate stronger perceived performance. The results show that our system is rated highest for analytical depth, integration of information, and support for humanitarian decision-making, while SmartBook performs strongly on ease of use. The summarizer receives predominantly neutral ratings due to its lack of structure and citations.

"Strongly Agree" responses. SmartBook performs well on ease-of-use and workflow integration, reflected in the higher scores shown in Figure 3 for these specific dimensions. The expert evaluation demonstrates that our system performs comparably to SmartBook on several core dimensions of report quality, with both systems sharing a similar overall approach while differing in specific implementation details. However, these advantages must be interpreted in light of its TF-IDF–based clustering, which is well suited to news-like corpora but less compatible with heterogeneous humanitarian documents like ours. This approach led to many clusters containing only two documents, which SmartBook discards, resulting in the exclusion of a substantial portion of the corpus and, consequently, the generation of shorter and simpler reports. This reduction in coverage explains the high usability ratings but also highlights an important limitation when handling heterogeneous humanitarian text.

Reliability is an area where none of the systems received particularly strong evaluations. This outcome is consistent with the stringent expectations of the humanitarian domain, where even a single incorrect or unsupported citation can undermine trust in the entire report. To contextualize the reliability scores, we evaluated a sample of 150 claims generated by SmartBook; The results closely mirrored those obtained for our system: 81% of citations were relevant or partially relevant and 84% of claims were supported or partially supported. Notably, however, SmartBook achieved these metrics on a truncated document set; our RAG and post-processing pipeline likely offers a more reliable solution for accurate citation alignment when processing the full volume of evidence. This suggests that, under equivalent evidence conditions, our approach is likely to provide more robust citation alignment. The summarizer system, while faster, retains the advantage of grounding all content strictly in the source texts but lacks the organizational depth and analytical features that experts deem essential for complex situational analysis.

In terms of informativeness and coverage, both our system and SmartBook outperform the baseline summarizer, which received consistently neutral ratings owing to its lack of citations and organizational structure. When asked to provide an overall ranking, three of the four experts selected our system as their preferred choice, citing its greater comprehensiveness and navigability relative to the alternatives. They attributed these advantages to the integration of multiple organizational modes (subtopics and SDGs), granular question–answer structures, and higher-level summaries. These features were viewed as essential for synthesizing complex humanitarian crises. In the final ranking task, three of the four experts selected our system as their preferred option. The qualitative feedback indicates that while no single visualization mode is universally optimal, the ability to switch between complementary organizational structures, together with comprehensive evidence integration and transparent citation, makes the system particularly well suited for operational use. One of the experts further noted that they did not regard the SDG-based organization as a foundational structure for situational reporting, being more accustomed to sector- or cluster-based categorizations (such as WASH, health, or protection), while still valuing the SDG view as a complementary

ARTICLE IN PRESS

lens alongside the sub-topic organization.

Finally, we note that the evaluations were conducted across several distinct humanitarian events. This diversity likely contributed to some variability observed in the scores for information quality and report usefulness, reflecting differences in the quality and density of source materials. Similar variability was previously observed in the question generation stage and is consistent with differences in document availability and specificity across events.

Computational Efficiency

To assess the computational efficiency of the proposed framework, we compared the total processing time required to generate situational reports across the three systems on a MacBook Pro (13-inch, 2021) equipped with a 2.3 GHz Quad-Core Intel Core i7 processor and 32 GB of RAM. This corresponds to a standard consumer laptop without dedicated GPU acceleration, so the reported times are an indicative benchmark rather than a hardware requirement; because the most demanding step—large-language-model inference—runs remotely through cloud APIs, a less powerful machine would mainly lengthen the local stages (embedding, clustering, and retrieval) rather than preclude use. As expected, the large-context summarizer was the fastest, owing to its comparatively simple architecture, which performs direct multi-document summarization without intermediate analytical steps. Processing times for our system and SmartBook were broadly comparable. On average, our pipeline required an additional 2–3 minutes to complete a report. This difference is attributable to components unique to our design, including the generation of executive summaries, question filtering, RAG-based answer extraction with post-processing of citations, and the production of cluster-level and SDG-level summaries. While these steps introduce modest computational overhead, they substantially enhance report depth, structure, and analytical utility.

From an operational cost perspective, only components relying on GPT-4o incur per-execution expenses, while the majority of models employed, such as Gemini 2.5 Flash and ModernBERT, are freely available. Despite the inclusion of GPT-4o, the average cost of generating a full report remains below $2.00 USD per event. Notably, for several tasks (e.g., question answering), the freely available Gemini 2.5 Flash model not only reduced costs but also produced higher-quality outputs in preliminary evaluations.

Importantly, the modular design of the framework enables flexible substitution of model components, making it straightforward to optimize for speed, cost, or performance as needed. Further runtime reductions could be achieved by batch retrieval strategies, parallelization of cluster-level operations, or integration of smaller, domain-adapted generative models. Benchmark execution times for three representative events (marked with * in Table 1a) are provided in the Supplementary Materials.

Discussion

This study demonstrates that large language models can effectively support the automatic generation of humanitarian situation reports, one of the most demanding analytical tasks in crisis response[4]. The proposed framework integrates document clustering, automatic question generation, retrieval-augmented answer synthesis, and executive summary generation within a coherent pipeline designed to approximate the reasoning workflow of human analysts. Applied across 13 humanitarian events and more than 1,100 source documents, the system successfully extracted, organized, and synthesized critical information with high factual accuracy and consistency. In this sense, the framework is intended to change how analysts access evidence—surfacing and linking the information relevant to a decision so it can be reached and traced directly—rather than to increase how often full reports are read.

Five humanitarian experts with direct experience producing and interpreting situational reports evaluated the system. They highlighted several strengths. The generated questions were consistently rated as highly relevant, important, and urgent, and the corresponding answers were judged to be well grounded in evidence with generally precise citation alignment. Experts from UNICEF further emphasized the system's operational usefulness and its potential for integration into existing analytical workflows. Compared with existing approaches, most experts preferred our system, citing the flexibility provided by its multiple organizational modes (subtopics and SDGs), the usefulness of executive summaries, and its capacity to structure and contextualize information in a manner aligned with established humanitarian analytical practices. These features collectively enhanced usability and analytical depth relative to baseline systems. Although SmartBook processed some document types more quickly, our system's additional components, including executive summaries, question filtering, and cluster-level synthesis, yielded reports that were more comprehensive and better adapted to heterogeneous humanitarian text sources. Collectively, these findings suggest that the framework can substantially reduce the effort required to synthesize large volumes of humanitarian information while maintaining transparency through source-linked evidence.

As with any evaluation of a complex AI system in a high-stakes domain, these findings should be interpreted in light of several important limitations. First, the quality and representativeness of the generated reports depend directly on the underlying document corpus. Variability in source quality, reporting density, and geographic coverage propagates through later stages of the pipeline and can affect both the specificity of generated questions and the informativeness of synthesized answers. Moreover, the framework inherits a selection bias from its input corpus: the composition of the source platforms reflects upstream decisions

ARTICLE IN PRESS

about which crises, regions, and languages are documented, and the resulting reports inevitably mirror this coverage. As the pipeline is source-agnostic, its representativeness depends directly on the breadth of the documents provided; the corpus used here therefore bounds the generality of our findings, and coverage can be broadened by incorporating additional or community-sourced data flows. Although the framework aims to reduce the effort required to access and synthesize information, we did not evaluate whether it reduces information overload or increases report consumption among intended users. Accordingly, our expert evaluation assessed report quality, navigability, transparency, and overall usefulness rather than report readership or changes in information-consumption behavior. Understanding how automated reporting affects information-seeking behavior represents an important future research direction. A related, longer-term concern is that automated generation lowers the cost of producing such reports at scale: if widely adopted and disseminated, these and similar machine-generated outputs could increase the share of synthetic content in the document ecosystem from which retrieval corpora are drawn, allowing biases to propagate and become progressively harder to detect over time. We did not evaluate this effect, which lies beyond the present study and represents an important direction for future work. Potential mitigation strategies include restricting retrieval to trusted repositories, tracking document provenance, and periodically auditing corpora for machine-generated content, although the effectiveness of such measures remains to be systematically evaluated.

Citation reliability, while generally strong, is not perfect: occasional citation errors persist, and even a single unsupported or incorrect citation can undermine trust in operational settings where precise attribution is essential. These errors are not uniform in severity: unnecessary (4.3%) and inferred citations (4.5%) mainly degrade transparency, whereas the incorrect citations (5.1%) and hallucinations (3.0%) pose genuine factual risks—a distinction that matters for a system meant to assist rather than replace human analysts. Quantitative claims are especially hard, as casualty, displacement, and funding figures change rapidly and are reported inconsistently across sources, so the same event can yield conflicting numbers that no source-grounded system can fully reconcile. These findings should not be interpreted as evidence that citation verification can be fully automated in high-stakes humanitarian settings. Rather, they suggest that the framework can substantially reduce the effort required for information synthesis while still benefiting from human verification of critical claims, numerical estimates, and attribution-sensitive statements. More broadly, the system currently processes only textual inputs, excluding potentially valuable information contained in images, tables, and geospatial data. Runtime costs also constrain real-time or large-scale deployment, though the modular architecture facilitates future optimization and substitution with more efficient or domain-adapted models and the same modularity offers a direct path to improving reliability: individual components-retrieval, citation post-processing, or the LLM-as-a-judge layer-can be upgraded or replaced as more accurate methods emerge, without redesigning the pipeline. While experts judged the generated reports to be more useful, navigable, and transparently sourced than the comparison systems, our evaluation focused on perceived report quality rather than downstream decision outcomes. We therefore do not claim that these improvements necessarily translate into better humanitarian decisions, coordination, or resource allocation. Assessing the operational impact of AI-assisted reporting through realistic decision-making exercises or longitudinal field deployments remains an important direction for future work. Finally, our evaluation rests on a small panel of five experts with a predominantly analytical and data-science profile. Although their judgments are domain-relevant and informed by professional experience, the sample is not statistically representative of the broader humanitarian community, which includes field responders, cluster coordinators, operational managers, and policy specialists. The findings should therefore be interpreted as evidence of expert-informed usefulness rather than a definitive assessment of community-wide preferences.

Future work should address these limitations by extending the framework to multimodal and multilingual inputs, leveraging vision–language models to incorporate visual and structured evidence, thereby capturing a richer representation of humanitarian conditions. Additional factuality metrics, such as FactScore[66], may further strengthen robustness, while systematic evaluation across general, specific, and quantitative question types could clarify the system's strengths and failure modes. The framework currently generates all questions dynamically; combining these with a fixed set of predefined questions—for instance, recurring items an organization always tracks—would be a straightforward extension that improves consistency across reports while preserving context-specific coverage. We also did not evaluate cross-event consistency in the information surfaced by dynamically generated questions. Future work should investigate hybrid approaches that combine fixed organizational reporting requirements with dynamically generated questions tailored to the specific crisis context. Embedding human feedback throughout the pipeline will remain essential for ensuring oversight and accountability as such systems move toward operational integration. A natural next step is a targeted human-in-the-loop stage in which domain experts validate high-stakes claims using the paragraph-level citations the system already provides, raising precision and reducing incorrect citations. Experts also recommended enabling more granular control over report length, beyond fixed presets, and incorporating additional visualization features, including timelines, severity indicators, and cluster-level summaries, to support rapid situational understanding. Linking the framework with existing platforms such as HDX Signal could enhance real-time monitoring and longitudinal crisis analysis.

A complementary finding of this study concerns the emerging role of LLMs as evaluators. The strong correspondence observed between human judgments and LLM-based assessments suggests that automated evaluation, when properly guided,

ARTICLE IN PRESS

can serve as a credible and scalable proxy for expert review[17,48]. Although LLM-based judgments cannot fully replicate human reasoning, their consistency indicates potential for more auditable and self-monitoring AI systems in humanitarian analysis.

In conclusion, the results provide empirical evidence that large language models, when coupled with structured retrieval and evaluation mechanisms, can substantially support the generation of accurate, verifiable, and interpretable humanitarian situation reports while indicating that full reliability—particularly for quantitative content and high-stakes attribution—remains an open challenge that requires sustained human oversight. To support transparency and reproducibility, we render the code and prompts used in this work publicly available, while the dataset is available upon request (see Data Availability). By bridging the gap between information extraction and operational decision-making, the proposed framework advances both the scientific and practical understanding of how generative AI can support humanitarian action at scale. Incorporating continued expert feedback on content prioritization, structural flexibility, visualization features, and usability will be essential for developing next-generation systems capable of supporting deployment in complex crisis environments.

Data Availability

The dataset used in this study was provided by Data Friendly Space (DFS) under a Memorandum of Understanding and includes copyrighted material from third-party humanitarian information platforms and news outlets. Due to licensing restrictions, the raw documents, annotations, and model outputs cannot be publicly released. Researchers interested in accessing the dataset may contact the corresponding authors, who will facilitate communication with DFS and support access requests in accordance with the data provider's governance policy.

The source code, prompt templates, analysis scripts, and the final generated reports are openly available in the project repository at https://github.com/idecost/LLM-SituationalReports. Although the full texts cannot be shared due to licensing, links to the original, publicly viewable source documents are provided within the repository. The generated reports can also be visualized via the online viewer at https://idecost.github.io/LLM-SituationalReports/Viewer/viewer_v2.html.

Acknowledgments

The authors acknowledge support from the Lagrange Project of the ISI Foundation, funded by Fondazione CRT. K.K. thanks AECID (Spanish Agency for International Development Cooperation) for their support to data innovation and Frontier Data Technologies through UNICEF's Frontier Data Network.

Author contributions

I.D., Y.M., and K.K. jointly conceived and designed the study. I.D. developed the codebase, implemented the framework, and conducted the data analysis. All authors contributed to interpreting the results, drafting sections of the manuscript, and revising it for important intellectual content. Y.M. and K.K. supervised the project and provided methodological and domain expertise throughout the study. All authors reviewed and approved the final manuscript.

References

1. New South Wales Rural Fire Service. *Situation Reports*. NSW Rural Fire Service (2024). Operational Guideline, Foundational.
2. of Sciences, N. A. *et al. A decadal survey of the social and behavioral sciences: A research agenda for advancing intelligence analysis* (National Academies Press, 2019).
3. IFP News. Un report confirms nobody reads un reports (2024). Accessed: 2025-08-17.
4. ALNAP (Active Learning Network for Accountability and Performance). Explain: Essential briefings for humanitarian decisionâ€‘makers â€“ artificial intelligence in the humanitarian sector. PDF briefing, ALNAP (2025). Longâ€‘read version.
5. CHS Alliance, Groupe URD and Sphere. The Core Humanitarian Standard on Quality and Accountability. https://corehumanitarianstandard.org/ (2024). Accessed: 2026-06-10.
6. Cao, H. Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark. *arXiv preprint arXiv:2406.01607* (2024).
7. Hadi, M. U. *et al.* Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea preprints* **1**, 1–26 (2023).

ARTICLE IN PRESS

8. Abdi, A., Idris, N., Alguliyev, R. M. & Aliguliyev, R. M. Query-based multi-documents summarization using linguistic knowledge and content word expansion. *Soft Comput.* **21**, 1785–1801 (2017).
9. Fekih, S. *et al.* Humset: Dataset of multilingual information extraction and classification for humanitarian crises response. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4379–4389 (2022).
10. Data Friendly Space. GANNET: AI-powered humanitarian intelligence platform. https://gannet.ai/ (2025). Accessed: 2026-06-10.
11. Bao, S., Li, T. & Cao, B. Chain-of-event prompting for multi-document summarization by large language models. *Int. J. Web Inf. Syst.* **20**, 229–247 (2024).
12. Reddy, R. G. *et al.* Smartbook: Ai-assisted situation report generation. *arXiv preprint arXiv:2303.14337* (2023).
13. Thulke, D. *et al.* Climategpt: Towards ai synthesizing interdisciplinary research on climate change. arxiv. 2024 doi: 10.48550. *arXiv preprint arXiv.2401.09646* (2024).
14. Ceresa, M. *et al.* Retrieval augmented generation evaluation for health documents. *arXiv preprint arXiv:2505.04680* (2025).
15. Sai, A. B., Mohankumar, A. K. & Khapra, M. M. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv. (CSUR)* **55**, 1–39 (2022).
16. Gu, J. *et al.* A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).
17. Li, H. *et al.* Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).
18. Stevens, C. & Kanie, N. The transformative potential of the sustainable development goals (sdgs). *Int. Environ. Agreements: Polit. Law Econ.* **16**, 393–396 (2016).
19. United Nations, Department of Economic and Social Affairs. The 17 Sustainable Development Goals. https://sdgs.un.org/goals (2025).
20. Gorgun, G. & Bulut, O. Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Educ. Inf. Technol.* **29**, 24111–24142, DOI: 10.1007/s10639-024-12771-3 (2024).
21. Hasan, A., Ehsan, M. A., Shahnoor, K. B. & Tasneem, S. S. *Automatic question & answer generation using generative large language model (llm)*. Ph.D. thesis, Brac University (2024).
22. UN Office for the Coordination of Humanitarian Affairs (OCHA). Situation Report (Sitrep). OCHA Information Management Toolbox, https://knowledge.base.unocha.org/wiki/spaces/imtoolbox/pages/42047193/Situation+Report+Sitrep (2023). Accessed: 2026-06-10.
23. CARE International. Situation Reports (Sitreps). CARE Emergency Toolkit, https://www.careemergencytoolkit.org/meal/42-information-management/4-situation-reports-sitreps/ (2018). Accessed: 2026-06-10.
24. Bavaresco, A. *et al.* Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403* (2024).
25. McInnes, L., Healy, J., Astels, S. *et al.* hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).
26. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. neural information processing systems* **33**, 9459–9474 (2020).
27. Ahmed, M. H., Tiun, S., Omar, N. & Sani, N. S. Short text clustering algorithms, application and challenges: A survey. *Appl. Sci.* **13**, 342 (2022).
28. Asyaky, M. S. & Mandala, R. Improving the performance of hdbscan on short text clustering by using word embedding and umap. In *2021 8th international conference on advanced informatics: Concepts, theory and applications (ICAICTA)*, 1–6 (IEEE, 2021).
29. Warner, B. *et al.* Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663* (2024).
30. Durrance-Bagale, A. *et al.* Lessons from humanitarian clusters to strengthen health system responses to mass displacement in low and middle-income countries: a scoping review. *J. Migr. Heal.* **1–2**, 100028 (2020).
31. Nussbaum, Z., Morris, J. X., Duderstadt, B. & Mulyar, A. Nomic embed: Training a reproducible long context text embedder (2024). 2402.01613.

ARTICLE IN PRESS

32. Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

33. Muennighoff, N., Tazi, N., Magne, L. & Reimers, N. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316* (2022).

34. Raunak, V., Gupta, V. & Metze, F. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 235–243 (2019).

35. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

36. Allaoui, M., Kherfi, M. L. & Cheriet, A. Considerably improving clustering algorithms using umap dimensionality reduction technique: a comparative study. In *International conference on image and signal processing*, 317–325 (Springer, 2020).

37. Saxena, A. *et al.* A review of clustering techniques and developments. *Neurocomputing* **267**, 664–681 (2017).

38. Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A. & Sander, J. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, 839–847 (SIAM, 2014).

39. Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S. & Hupkes, D. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624* (2024).

40. Vysala, A. & Gomes, J. Evaluating and validating cluster results. *arXiv preprint arXiv:2007.08034* (2020).

41. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. neural information processing systems* **35**, 24824–24837 (2022).

42. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).

43. Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

44. Sultan, M. A., Chandel, S., Astudillo, R. F. & Castelli, V. On the importance of diversity in question generation for qa. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5651–5656 (2020).

45. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).

46. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

47. Comanici, G. *et al.* Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

48. Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **36**, 46595–46623 (2023).

49. Wataoka, K., Takahashi, T. & Ri, R. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819* (2024).

50. Li, S., Stenzel, L., Eickhoff, C. & Bahrainian, S. A. Enhancing retrieval-augmented generation: A study of best practices. *arXiv preprint arXiv:2501.07391* (2025).

51. Fan, W. *et al.* A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491–6501 (2024).

52. Qian, H., Fan, Y., Zhang, R. & Guo, J. On the capacity of citation generation by large language models. In *China Conference on Information Retrieval*, 109–123 (Springer, 2024).

53. Slobodkin, A., Hirsch, E., Cattan, A., Schuster, T. & Dagan, I. Attribute first, then generate: Locally-attributable grounded text generation. *arXiv preprint arXiv:2403.17104* (2024).

54. Khattab, O. & Zaharia, M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 39–48 (2020).

55. Rackauckas, Z. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367* (2024).

56. Chiang, W.-L. *et al.* Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning* (2024).

57. Maheshwari, H., Tenneti, S. & Nakkiran, A. Citefix: Enhancing rag accuracy through post-processing citation correction. *arXiv preprint arXiv:2504.15629* (2025).

ARTICLE IN PRESS

ARTICLE IN PRESS

58. Cooper, W. S. A definition of relevance for information retrieval. *Inf. storage retrieval* **7**, 19–37 (1971).

59. Cohen, J. A coefficient of agreement for nominal scales. *Educ. psychological measurement* **20**, 37–46 (1960).

60. Byrt, T., Bishop, J. & Carlin, J. B. Bias, prevalence and kappa. *J. clinical epidemiology* **46**, 423–429 (1993).

61. Delgado, R. & Tibau, X.-A. Why cohenâ€™s kappa should be avoided as performance measure in classification. *PloS one* **14**, e0222916 (2019).

62. Es, S., James, J., Anke, L. E. & Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158 (2024).

63. Gao, T., Yen, H., Yu, J. & Chen, D. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627* (2023).

64. Belliardo, E., Kalimeri, K. & Mejova, Y. Leave no place behind: improved geolocation in humanitarian documents. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, 31–39 (2023).

65. An, B., Zhang, S. & Dredze, M. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. *arXiv preprint arXiv:2504.18041* (2025).

66. Min, S. *et al.* Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100 (2023).

Competing interests

The author(s) declare no competing interests.

Funding

This study was partially funded by Fondazione CRT through the Lagrange Project of the ISI Foundation. K.K. additionally received support from the Spanish Agency for International Development Cooperation (AECID) via UNICEF's Frontier Data Network.

ARTICLE IN PRESS