Disaster Response Report Generation System

Ming-Lu Liu[1] and James Yichu Chen[1]

[1]Department of Civil and Construction Engineering, National Taiwan University of Science and Technology, Taiwan
m11205504@gapps.ntust.edu.tw, yie@mail.ntust.edu.tw

Abstract –

This study develops an automated disaster response report generation system that integrates large language models (LLMs), web scraping, and computer vision techniques. The system automatically retrieves data from 16 heterogeneous and distributed sources and, by combining Retrieval-Augmented Generation (RAG) with Chain-of-Thought (CoT) prompting, produces first drafts of disaster response reports that conform to the official reporting framework. Empirical results show that the system achieves a System Usability Scale (SUS) score of 85.8 and reduces report preparation time by approximately 73% for typhoon events and 64% for heavy-rain events. Expert evaluations indicate that the generated content attains acceptable levels of completeness, accuracy, and readability. To further validate content reliability, TLM trustworthiness scores of 0.89 and 0.95 confirm high factual reliability, while RAGAS Faithfulness scores of 0.99 (RAG–CoT) versus 0.43 (vanilla LLM) demonstrate the contribution of the proposed framework. These findings confirm the feasibility and practical potential of the proposed system for automated report generation in professional disaster management contexts.

Keywords –

Disaster Response, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT), Credibility Evaluation

1 Introduction

In recent years, extreme weather events have become increasingly frequent worldwide, and the associated economic and social impacts of natural disasters continue to escalate. In Taiwan, weather-related disasters cause average annual economic losses of approximately NT$15 billion (about US$490 million). According to statistics from the Water Resources Agency (WRA), the number of disaster response operations has increased markedly, from about 40–50 per year between 2015 and 2020 to 60–90 per year after 2020. The growing frequency of response operations has substantially increased the workload of frontline personnel. In particular, the preparation of post-disaster response reports—which must be completed within 30 days and serve as critical records for institutional learning—has become one of the most labor-intensive components of disaster response work.

At present, disaster response reports are still written largely by hand through a complex and time-consuming process. Taking the 2024 Typhoon Gaemi response report as an example, the preparation of this report required integrating 16 distinct data sources, compiling 11 tables, and producing 26 figures, resulting in a document of 163 pages. The entire process consumed approximately three full working days. Moreover, the sections describing meteorological evolution require substantial domain expertise; new staff members often need several months of training before independently drafting such reports, and during peak flood seasons, report writing must be conducted concurrently with ongoing emergency operations, further intensifying the burden on already-limited personnel.

In response to these challenges, this study proposes an automated disaster response report generation system that integrates LLMs, web scraping, and computer vision. The system primarily targets typhoon events and large-scale heavy-rain events, and seeks to improve report-writing efficiency through automated data acquisition, chart generation, and draft text generation. Furthermore, the system adopts a RAG–CoT framework and incorporates TLM-based evaluation to achieve the following objectives:

1. Reduce workload and improve efficiency by automating the labor-intensive processes of data collection and initial report drafting.
2. Integrate scattered heterogeneous data into standardized reporting templates, thereby lowering the technical threshold and facilitating the rapid onboarding of new staff.
3. Validate generative performance and content reliability through a hybrid evaluation framework combining expert assessment, TLM credibility scoring, and RAGAS faithfulness metrics.

2 Related Work

This section reviews literature relevant to the proposed system. First, it introduces the structure and

data integration challenges of disaster response reports in Taiwan. Second, it examines the automated report generation technologies adopted in this study, including web scraping, computer vision, and LLMs, and discusses hallucination issues and corresponding mitigation strategies. Finally, it surveys the current applications of LLMs in disaster response, which form the theoretical foundation of this research.

2.1 Disaster Response Report Standards and Architecture

Disaster response reports are key documents in the post-disaster recovery and mitigation cycle. In Taiwan, the Water Resources Administration of the Ministry of Economic Affairs has formulated structured writing standards for disaster response reports, requiring them to systematically present the background of the incident, meteorological and hydrological evolution, response operations, and the actual impact of disasters. These reports cover three core modules: event background and meteorological developments, hydrological status and disaster impact, and response operations and notification records.

However, producing such standardized disaster response reports relies heavily on cross-system data integration. In practice, the information required for the reports is distributed across multiple independent information systems that differ substantially in technical architecture, data formats, and access mechanisms. This study identifies 16 distinct data sources dispersed across 6 web platforms (Table 1), making cross-system integration a highly complex technical challenge. These data sources are maintained by various government agencies and include HTML tables, JSON data, image files (JPG/PNG), and unstructured text. Issues such as missing data and inconsistent time stamps are common. Under tight response timelines, specialists must repeatedly access multiple systems to retrieve and reconcile data, highlighting both the cumbersome nature of existing workflows and the urgent need for automation.

Table 1 Classification of data sources

Category	Primary Source Platforms and Formats
Water Conservancy Monitoring	Alert Issuance Log, Disaster Emergency Response System (TXT, HTML, JSON, DOCX)
Meteorological Information	Weather Analysis and Taiwan Climate Hybrid monitor system (WATCH), Typhoon Database (TXT, JPG, PNG)
External Warning	Public Open Data Platform (TXT)
Marine Conditions	Safe Ocean (XLSX)

2.2 Report Automation and LLMs-Based Technology

In the field of disaster management, automation has been widely adopted to handle multi-source heterogeneous data and real-time requirements. With respect to data collection, web scraping combined with API access has been used to automatically aggregate disaster-related information from government open data portals, alert platforms, and news or social media streams, thereby improving situational awareness[1, 2]. However, most of these studies focus on real-time monitoring dashboards rather than on the long-term event histories and multi-chapter structures required for formal reporting.

For data visualization and chart understanding, prior work has shown that directly summarizing charts using end-to-end language models can easily lead to numerical hallucinations[3]. A more reliable approach is to first recover intermediate structured data (e.g., tables or statistical matrices) from chart images and then apply data-to-text generation, which improves the consistency between textual descriptions and the underlying quantitative data[4]. This suggests that, in the context of formal disaster reports, converting meteorological and hydrological maps into machine-readable structured indicators is an important prerequisite for trustworthy text generation.

LLMs have been widely applied in text summarization, question answering, and automated report generation. Nevertheless, in high-risk professional domains such as disaster response, content accuracy and hallucinations remain major concerns[5, 6]. To address these issues, recent research has proposed strategies such as RAG and CoT[5]. RAG dynamically retrieves verifiable external information during generation, grounding model outputs in authoritative documents and reducing hallucination rates. CoT prompting encourages models to perform step-by-step reasoning, decomposing complex tasks into a sequence of intermediate reasoning steps, thereby improving the structure and interpretability of responses. Empirical studies have shown that combining RAG and CoT can enhance the reliability of generated professional texts while maintaining overall performance.

In the disaster response domain, automated reporting pipelines and disaster information systems have been developed to support situation awareness and documentation, while recent studies survey broader AI applications across multiple phases of disaster management, from early warning to post-event analysis. Recent interdisciplinary reviews indicate that AI and

LLM applications in disaster management span multiple phases—from detection and situation analysis to decision support and post-event documentation—with social media analysis and decision recommendation representing the most mature application areas[7]. For instance, DisasterResponseGPT integrates response guidelines and expert knowledge into prompts to accelerate the formulation of response plans[8]. Beyond individual LLM applications, knowledge graph-enhanced systems have been developed to provide evidence-based recommendations in emergency decision-making scenarios[9], and multi-agent AI frameworks have been proposed to integrate sensing, analysis, and communication functions into unified disaster management architectures [10]. Other studies employ LLMs to aggregate disaster-related texts and integrate multi-source information[11, 12]. However, existing approaches rarely integrate visual information, structured operational data, and standardized governmental reporting formats, and few quantitatively assess content credibility. This study addresses these gaps by proposing a RAG-CoT framework that automatically generates structured, traceable first drafts of Taiwan's water-related disaster response reports with quantitative credibility evaluation..

3 Methodology

This section describes the architecture and implementation of the proposed disaster response report generation system. The core design principle is human–AI collaboration: automated components handle highly repetitive data-integration tasks, while LLMs assist with drafting narrative sections, allowing domain experts to concentrate on high-value judgment and decision-making. The system supports the automatic generation of six sections of a complete disaster response report: Event Overview, Meteorological Situation, Hydrological Information, Emergency Response Team Operations, Disaster Impacts, and Conclusion and Annexes. Among these, the first three sections are highly narrative and require the integration of a large volume of meteorological and hydrological information; consequently, they are the primary focus of the LLMs-based generation strategies in this study.

3.1 System Architecture

The overall system architecture is shown in **Error! Reference source not found.**. It consists of six main modules: Data Collection, Data Acquisition, Data Visualization, Data Analysis, Text Generation, and Report Integration. Through the user interface, report compilers specify the event name and the time ranges for rainfall and water-level data. The Data Collection and Data Acquisition modules then automatically retrieve the raw data required for each chapter from multiple platforms. These data are passed to the Data Visualization and Data Analysis modules, which generate the figures and analytical results required in the report. The Text Generation module uses the structured datasets and analytical outputs to produce draft text for the corresponding sections. Finally, the Report Integration module assembles all generated texts, tables, and figures into a pre-defined Word template, thereby producing a first draft of the disaster response report for the selected event.

3.2 Clarification of Data Sources and Processing Pipelines

To ensure that the automated system can effectively support real-world reporting requirements, the content of the disaster response report was first analyzed and categorized according to information type and generation risk. The report includes both highly descriptive narrative texts and quantitative content—tables and figures—for which numerical accuracy is critical. Relying on a single generation strategy for all content would likely yield uneven quality and factual errors. Accordingly, this study adopts a diverted generation design, in which different sections are handled by different technical pipelines depending on their semantic and numerical requirements. The narrative sections—Event Overview, Meteorological Situation, and Hydrological Information—must synthesize a large number of weather bulletins, observational records, and response logs into coherent textual descriptions. For these sections, the system leverages LLMs with integrated RAG and CoT prompting. Automatically collected and pre-processed meteorological and hydrological data are converted into reference documents that serve as contextual inputs for the LLMs, which in turn generates draft paragraphs aligned with official reporting style and structure. In contrast, sections such as the charts in Hydrological Information, Emergency Response Team Operations, Disaster Impacts, Conclusion, and Annexes are primarily composed of numerical data, statistical tables, and visualizations. The key objectives for these sections are numerical correctness and consistent formatting rather than free-form narrative. For them, the system relies on automated data processing and visualization pipelines to perform data cleaning, computation, and figure/table generation. In practice, the pipeline consists of three steps: heterogeneous inputs from the 16 data sources are first converted into a unified, event-based schema with basic cleaning (e.g., missing-value handling and time alignment). The cleaned data are then aggregated into chapter-specific statistical indicators, which the visualization module uses to automatically generate standardized charts and tables and insert them into the predefined Word report template. This approach yields standardized charts and tables that conform to reporting specifications and ensures the reliability of the

quantitative content.

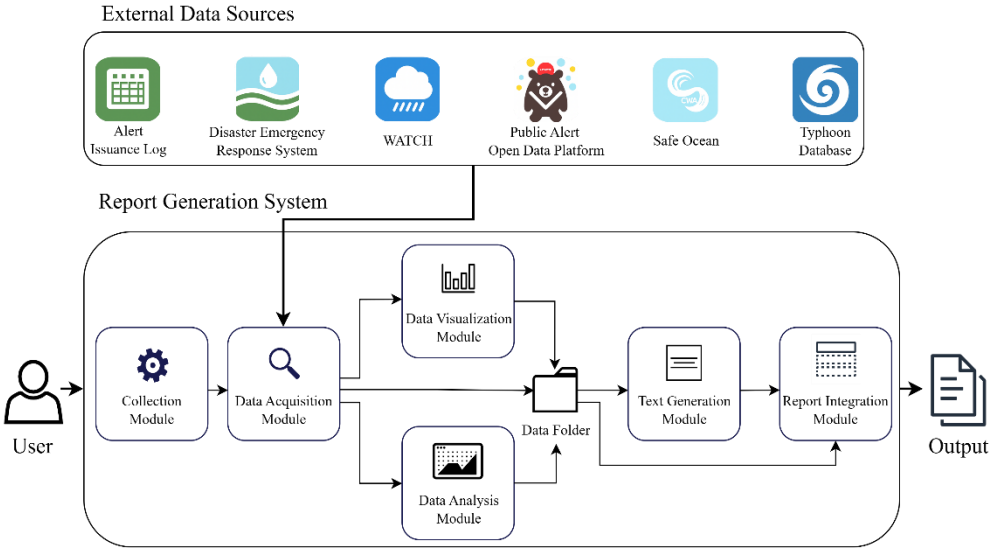

Figure 1 System architecture of the disaster response report generation system

3.3 RAG and CoT Implementation

To improve text generation quality and reduce hallucination risks, the system integrates RAG and CoT strategies, as illustrated in Figure 2. The multi-level prompting mechanism is designed to generate descriptive paragraphs for the Event Overview, Meteorological Situation, and Hydrological Information sections. For the RAG component, long-duration typhoon events can accumulate a large volume of textual data that exceeds the input length limits of current LLMs. To address this, the system first segments multi-day weather bulletins and related texts into smaller chunks, then encodes them into vectors using an embedding model and stores them in a FAISS vector database. During generation, the system performs similarity search based on the current prompt and retrieves the most relevant text fragments, which are supplied as contextual evidence to the LLMs. For the CoT component, prompts explicitly specify the reasoning steps required to construct each paragraph, decomposing complex generation tasks into sequential sub-tasks. Few-shot exemplars are included to demonstrate the desired writing style and paragraph structure of disaster response reports. For example, in the "Typhoon Development History" subsection, the prompt guides the model to chronologically summarize the typhoon track, key time points (e.g., warning issuance, landfall, departure, and cancellation), and to integrate rainfall and wind information at each stage, thereby producing narratives with clear timelines and causal relationships. In addition, hard constraints are embedded in the prompts to restrict the model to using only information retrieved by the RAG component, thereby discouraging unsupported speculation. Through the joint use of RAG and CoT, the system generates interpretable and well-grounded draft texts based on multi-source data, which are then subject to expert review and refinement.

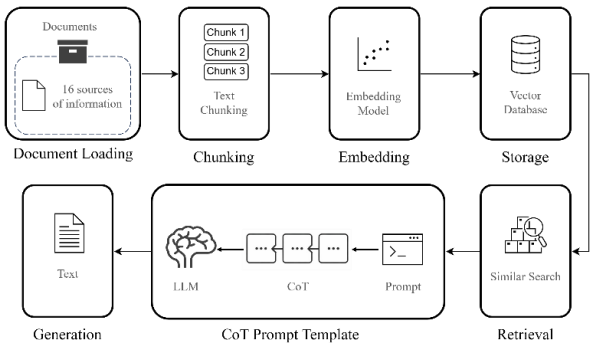

Figure 2 RAG - CoT architecture diagram

4 Validation

The system was evaluated across three stages: (1) system usability assessment, (2) expert evaluation of content quality and production efficiency, and (3) TLM credibility assessment. Test participants were personnel responsible for drafting disaster response reports at the Water Resources Administration of the Ministry of Economic Affairs of Taiwan.

4.1 Validation Methodology

Stage 1 – System usability.

The first stage examined whether the system effectively integrates dispersed data and reduces the

difficulty of report writing. Participants were asked to complete predefined tasks using the system without assistance. Their task completion time and operational difficulties were recorded. After the tasks, participants completed the SUS questionnaire and participated in semi-structured interviews to supplement the quantitative results. Six users with experience in disaster response report writing were recruited: two were classified as novice and four as experienced based on their years of practice, allowing the study to capture the perspectives of different user groups.

Stage 2 – Expert assessment of content quality and efficiency.

In the second stage, experts evaluated both the quality and efficiency of system-generated reports. One typhoon event and one heavy-rain event were selected as test cases, and experts with more than ten years of disaster response experience were invited to participate. For content quality, the experts rated each automatically generated paragraph on three dimensions—completeness, correctness, and readability—and compared the results with manually written versions to determine whether the generated content covered the necessary information and remained readable. All ratings used a five-point Likert scale. For production efficiency, experts recorded (i) the time required for the system to generate the first draft, (ii) the time needed to revise the drafts, and (iii) the time required to write the reports entirely by hand. The overall system-assisted production time was defined as the sum of generation and editing times and was compared with purely manual writing to calculate the percentage of time saved.

Stage 3 – Credibility assessment using TLM.

The third stage aimed to objectively and reproducibly evaluate the credibility of the system-generated texts, with particular attention to hallucination risks. Even when RAG is employed, LLMs may still produce misleading or incorrect statements; thus, an automated reliability check is necessary[13]. This study adopts the TLM proposed by Cleanlab[14] to evaluate the generated paragraphs in the Event Overview and Meteorological Situation and Hydrological Information sections. TLM quantifies the trustworthiness of responses based on the relationship among the user query, retrieved context, and generated output. The evaluation includes an overall trustworthiness score and four sub-indicators: context sufficiency, response groundedness, response helpfulness, and query ease. These metrics allow us to assess whether the generated content is sufficiently supported by retrieved evidence and internally consistent when synthesizing multi-source data, and the results also guide further system optimization. In addition, RAGAS is employed as a complementary evaluation tool to compare the RAG–CoT-enhanced configuration with a vanilla LLM using the same underlying model, quantifying differences in Faithfulness and Answer Relevancy and thereby demonstrating the contribution of the RAG–CoT architecture to overall content quality.

4.2 Sample Selection

To ensure representative samples for the Stage-3 TLM evaluation, disaster events that occurred in 2024 were screened using predefined criteria. Typhoon events were required to trigger at least a Level-1 response by the Ministry of Economic Affairs and to cause multiple flooding incidents across counties and cities. Large-scale heavy-rain events were required to trigger at least a Level-2 response or to result in 10 or more flooding locations. Based on these criteria, eight events were selected—four typhoon events and four heavy-rain events—as summarized in Table 2.

Table 2 List of selected typhoon and heavy-rain events

Event Name	Cumulative Rainfall (mm)	Screening Criteria
Typhoon GAEMI	1,944	Level 1 response: 3,308 floods in 15 counties and cities
Typhoon KRATHON	1,725	Level 1 response: 604 floods in 8 counties and cities
Typhoon KONG-REY	1,184	Level 1 response: 109 floods in 9 counties and cities
Typhoon USAGI	492	Level 1 response: 9 flooding in 3 counties and cities
0602 Heavy-Rain	223	Level 2 response: 12 floods in 5 counties and cities
0710 Heavy-Rain	135	Level 3 response: 55 flooding in 1 county and city
0806 Heavy-Rain	191	Level 3 response: 16 flooding in 2 counties and cities
0921 Heavy-Rain	436	Level 2 response: 83 floods in 6 counties and cities

4.3 Result

The SUS was used to evaluate overall system usability. According to Bangor et al. (2009), an average SUS score of 68 is generally considered the benchmark for acceptable usability[15, 16]. The proposed system achieved an average SUS score of 85.8 (Figure 3), which falls into the Excellent range, indicating a high level of usability in terms of both operation and user experience.

Further analysis revealed clear differences between novice and experienced users. Experienced staff scored the system between 92.5 and 100, reflecting strong recognition of its benefits for data integration and report preparation. Novice users, by contrast, assigned scores between 57.5 and 65, slightly below the benchmark. This suggests that users who have previously written large-scale disaster reports are more capable of perceiving the practical advantages of the system for multi-source data integration and draft generation, whereas novices tend to focus on interface intuitiveness and may have limited understanding of the complex disaster-response scenarios the system is designed to support. Nevertheless, during follow-up interviews, novice users generally agreed that the automatically generated Event Overview, Meteorological Situation, and Hydrological Information sections helped them quickly grasp the report structure and served as useful templates for writing.

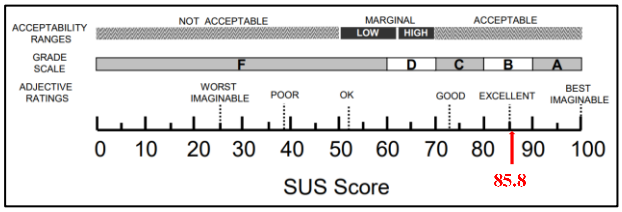

Figure 3 System Usability Scale score

Table 3 summarizes the expert evaluation results. For the specific content checklist, all chapters generated by the system obtained average scores of at least 3 out of 5, indicating that the model was able to cover the key information required in disaster response reports. For the overall quality dimensions—completeness, correctness, and readability—average scores were all at or above 4, suggesting that the generated content met the basic quality requirements of formal response reports. Experts noted that a small number of paragraphs could benefit from more precise wording and tighter narrative focus, but overall, only minor manual editing was needed before practical use. In terms of production efficiency, manually writing a typhoon event report required approximately 36 hours. With system assistance (automatic draft plus expert editing), the total time was reduced to about 9.60 hours, corresponding to a 73% reduction. For heavy-rain event reports, manual writing required about 2 hours, whereas the system-assisted process required approximately 0.73 hours, saving around 64% of the time. These results demonstrate that the proposed system can substantially improve report-writing efficiency for disaster events of varying scales.

TLM was employed to assess the credibility of the Event Overview and Meteorological Situation and Hydrological Information sections generated by the system. Scores range from 0 to 1; following Cleanlab's guidelines, values below 0.7 are considered low confidence, while values above 0.9 are regarded as high confidence. As shown in Table 4, the Event Overview and Meteorological Situation section achieved an average trustworthiness score of 0.89, close to the high-confidence range, whereas the Hydrological Information section achieved an average of 0.95, which clearly falls within the high-confidence range. These results indicate that the generated content is well supported by factual evidence and is overall reliable. Regarding the sub-indicators, average scores for context sufficiency, response groundedness, and response helpfulness all exceeded 0.9. In particular, response groundedness and response helpfulness both averaged 0.99, implying that the generated responses were highly dependent on retrieved evidence and exhibited virtually no hallucinations.

Table 3 Expert evaluation result table

Evaluation Dimension	Indicator	Evaluation Result
Completeness	Key Information Coverage	Average Score ≧ 3
Quality	Overall Quality Score	Average Score ≧ 4
Efficiency (Typhoon)	Report Generation Time	36 h → 9.60 h (73% Time Saved)
Efficiency (Heavy-Rain)	Report Generation Time	2 h → 0.73 h (64% Time Saved)

Slight differences were observed between disaster types: compared with typhoon events—where data continuity and publicly available information are more complete—heavy-rain events obtained a lower Query Ease score because the corresponding hydrological generation tasks must integrate heterogeneous evidence sources (textual alerts, numerical daily maximum rainfall data, and JSON-formatted rainfall-map analysis results) and follow prompts that require the model to infer daily weather patterns and affected areas during the response period, making these queries inherently more complex for the TLM to evaluate. Even so, both the Hydrological Information and Event Overview and Meteorological Situation sections remained at high credibility levels.

Overall, the TLM results demonstrate that the proposed system can consistently generate highly trustworthy and practically useful textual content in complex multi-source data settings.

Beyond assessing trustworthiness with TLM, this study further examined whether incorporating the proposed RAG–CoT framework improves the quality of LLM-generated reports. As shown in Table 5, both configurations achieved very similar Answer Relevancy scores (~0.82), indicating that the vanilla LLM is capable of generating topically relevant content. However, the large difference in Faithfulness (0.99 vs. 0.43) shows that vanilla LLM outputs frequently contain fabricated or unsupported statements, which is a critical risk in professional disaster reporting where factual accuracy is required. These results demonstrate that adopting the proposed RAG–CoT framework improves the factual grounding of automatically generated reports in this professional domain.

Table 4 Credibility assessment result table

Indicators/Chapters	Event Overview and Meteorological Situation	Hydrological Information
Trustworthiness	0.89	0.95
Context Sufficiency	0.98	0.93
Response Groundedness	0.99	0.99
Response Helpfulness	0.99	0.99
Query Ease	0.96	0.80

Table 5 vanilla LLM vs. RAG-CoT

Indicators/Configuration	Vanilla LLM	RAG–CoT	Improvement
Faithfulness	0.43	0.99	+ 0.56
Answer Relevancy	0.81	0.82	+ 0.01

5 Discussion

The disaster response report generation system developed in this study integrates web scraping, geospatial data analysis, and LLM-based text generation to automatically produce draft content for the Event Overview, Meteorological Situation, and Hydrological Information sections. Expert evaluations confirm that the generated content meets the basic quality requirements of formal disaster response reports, while the system reduces report preparation time by approximately 73% for typhoon events and 64% for heavy-rain events. Based on the 96 heavy-rain and 4 typhoon response events recorded in Taiwan in 2024, purely manual drafting would require an estimated 108 working days, whereas the proposed system reduces this workload to 17 working days, substantially easing the operational burden on response units during periods of intensive disaster activity. In addition, the SUS score of 85.8 (Excellent) indicates that the system achieves strong usability in practical settings.

Despite these promising results, several limitations remain. First, the system currently depends heavily on the availability and stability of external data sources and website structures; when source platforms change their interfaces or data formats, the scraping and parsing components must be updated accordingly. When critical data cannot be retrieved or are insufficient to generate a given section, the system automatically inserts a placeholder (e.g., “to be completed by responders”) into the draft report to clearly indicate that manual completion is required. Second, although the generated texts are generally factually correct, final review and refinement by domain experts are still required to ensure that the tone, wording preferences, and points of emphasis align with the conventions of specific agencies and with politically or socially sensitive issues. Third, in the case of small, localized heavy-rain events, the limited availability of official data can lead to insufficient context, causing the model to produce overly brief descriptions or explicit indications of missing data, which may reduce the perceived completeness and credibility of the final report.

In this study, the report structure is decomposed by analyzing each chapter’s content into three categories—narrative text, visual maps and charts, and numerical tables—and assigning each category to an appropriate technical module (LLM with RAG–CoT prompting for narrative sections, computer vision for map-derived indicators, and automated data processing and visualization pipelines for quantitative content). The same design principle can be applied in other regions or hazard types by re-analyzing the local report template,

classifying its required content into these three categories, and then adapting the data sources, computer-vision routines, and RAG document collections and prompt templates to local monitoring systems and reporting standards, while reusing the core RAG–CoT and report-integration architecture.

Future work will focus on three directions: developing an interactive query interface so that responders can use natural language to explore historical events and analogous cases, integrating historical reports and impact data to provide simple, context-aware response recommendations, and introducing AI agents and the Model Context Protocol (MCP) to enhance the system's ability to self-check, self-repair, and scale in real-world operational environments.

6 Conclusion

To address the practical need for responders to handle both real-time operations and formal report writing under high event frequency, this study proposes an automated disaster response report generation system that integrates LLMs, web scraping, and geospatial data analysis. The results show that, when coupled with a structured data-integration mechanism and interpretable generation strategies, LLMs can reliably support the production of first drafts of disaster response reports with acceptable credibility and practical utility. Overall, these findings confirm the practical feasibility of applying LLMs to governmental disaster-prevention documentation and public-sector reporting tasks, offering a modular and reusable framework adaptable to other professional domains.

Acknowledgement

This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant **NSTC 115-2625-M-011-001**. The authors also thank the domain experts and practitioners for their insightful feedback and contributions during system evaluation.

References

1. Kusumasari, B. and N.P.A. Prabowo, *Scraping social media data for disaster communication: how the pattern of Twitter users affects disasters in Asia and the Pacific.* Natural Hazards, 2020. **103**(3): p. 3415–3435.
2. Abid, S., et al. *Social Media Data: Challenges, Mitigation Strategies, and Opportunities for Disaster Management*. in *2024 International Conference on Frontiers of Information Technology (FIT)*. 2024.
3. Obeid, J. and E. Hoque, *Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model.* arXiv preprint arXiv:2010.09142, 2020.
4. Poco, J. and J. Heer. *Reverse-engineering visualizations: Recovering visual encodings from chart images*. in *Computer graphics forum*. 2017. Wiley Online Library.
5. Kumar, A., et al., *Improving the reliability of LLMs: Combining CoT, RAG, self-consistency, and self-verification.* arXiv preprint arXiv:2505.09031, 2025.
6. Naveed, H., et al., *A Comprehensive Overview of Large Language Models.* ACM Trans. Intell. Syst. Technol., 2025. **16**(5): p. Article 106.
7. Xu, F., et al., *Large language model applications in disaster management: an interdisciplinary review.* International Journal of Disaster Risk Reduction, 2025: p. 105642.
8. Goecks, V.G. and N.R. Waytowich, *Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios.* arXiv preprint arXiv:2306.17271, 2023.
9. Chen, M., et al., *Enhancing emergency decision-making with knowledge graphs and large language models.* International Journal of Disaster Risk Reduction, 2024. **113**: p. 104804.
10. Li, B., et al., *Disaster Management in the Era of Agentic AI Systems: A Vision for Collective Human-Machine Intelligence for Augmented Resilience.* arXiv preprint arXiv:2510.16034, 2025.
11. Colverd, G., et al., *Floodbrain: Flood disaster reporting by web-based retrieval augmented generation with an llm.* arXiv preprint arXiv:2311.02597, 2023.
12. Chen, Z., et al., *Integration of large vision language models for efficient post-disaster damage assessment and reporting.* arXiv preprint arXiv:2411.01511, 2024.
13. Sardana, A., *Real-Time Evaluation Models for RAG: Who Detects Hallucinations Best?* arXiv preprint arXiv:2503.21157, 2025.
14. Northcutt, C.G., *Overcoming Hallucinations with the Trustworthy Language Model.* 2024.
15. Bangor, A., P. Kortum, and J. Miller, *Determining what individual SUS scores mean: Adding an adjective rating scale.* Journal of usability studies, 2009. **4**(3): p. 114–123.
16. Lewis, J.R. and J. Sauro, *Item benchmarks for the system usability scale.* Journal of Usability studies, 2018. **13**(3).