More Than Rainfall: A Multi-Dimensional Analysis of LLM Flood Cause Multi-Document Summarization for Disaster Risk Management

Jane Arleth dela Cruz[1] **Jarno Van Arkel**[2] **Iris Hendrickx**[1] **Martha Larson**[1,2]
[1]Center for Language and Speech Technology, Center for Language Studies
[2]Institute for Computing and Information Sciences
Radboud University, Netherlands
{janearleth.delacruz, jarno.vanarkel, iris.hendrickx, martha.larson}@ru.nl

Abstract

Large language models (LLMs) are increasingly adopted in disaster risk management (DRM) for situational awareness tasks. An equally critical task is extracting and summarizing the underlying causes of a disaster for post-event reporting, which requires a higher-level of discourse comprehension. While pre-filtering the input data into "cause-specific" text seems to be a straightforward solution, cleaner inputs may not resolve all task-relevant issues. The limitations can stem from the LLMs' core synthesis behaviors. To investigate this, we conduct a two-part study on multi-document summarization of flood causes. Part I investigates the impact of input granularity (document-level vs. sentence-level input). Part II evaluates how LLM-generated summaries differ from human-written summaries using a targeted funnel evaluation, conducting a multi-dimensional human annotation across more than 20 variables. Our findings reveal that while the sentence-level extraction step reduces non-causal content for smaller models, advanced and larger LLMs did not show significant reductions. Furthermore, while LLMs match humans in baseline relevance, deeper synthesis divergences persist. Structurally, LLM-generated summaries are longer and less coherent. Semantically, LLMs over-generate extreme descriptors and under-represent the more nuanced higher-level governmental responsibility.

1 Introduction

Large language models (LLMs) have been adopted as easily deployable tools in the disaster risk management (DRM) domain, particularly in crisis informatics tasks (Lei et al., 2025; Xu et al., 2025) that support situational awareness during a crisis event. Existing crisis informatics work is largely focused on identifying and summarizing real-time impact information about disasters and prioritizing rescue efforts (McCreadie and Buntain, 2023; Lei et al., 2025; Imran et al., 2025). Multi-document summarization of online media data using LLMs can be useful for comprehensive reports (McCreadie and Buntain, 2023; Seeberger and Riedhammer, 2024). Yet, an equally important task is extracting and summarizing the underlying causes of a disaster event (Dela Cruz et al., 2026b). Cause-focused summaries can support accountability, mitigation planning, and retrospective reporting by synthesizing evidence about physical triggers, infrastructure conditions, human actors, and, where supported by source documents, institutional responsibility.

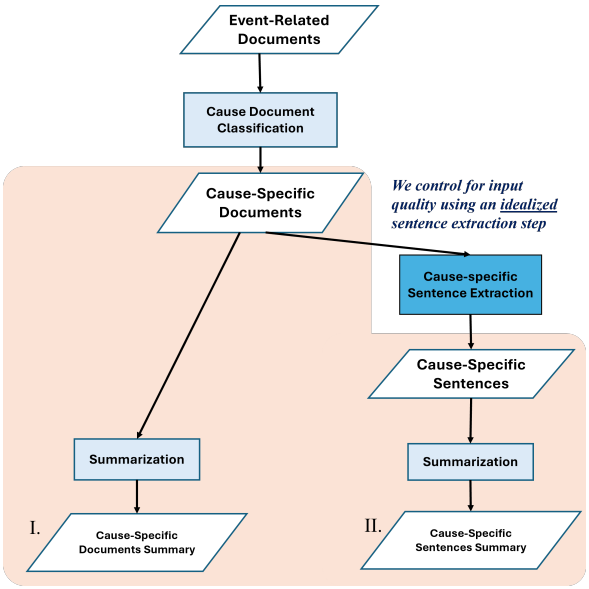


Figure 1: Two input conditions for flood cause summarization: I. Cause-specific documents (left) and II. Cause-specific sentences (right). Our investigation focuses exclusively on the final summarization step as highlighted in orange. We evaluate how input granularity affects the generated summaries given idealized, pre-filtered inputs. The upstream cause classification and extraction steps are out-of-scope in this work.

Flood cause summarization is a demanding form of query-based multi-document diversity summarization (Huang et al., 2024). Apart from recalling

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1074–1091
October 17-21, 2026 ©2026 Association for Computational Linguistics

event-related facts, a flood cause summary must capture diverse and interactive causal factors across documents, distinguish *why* the flooding occurred from *what* happened during the event, and organize these factors into a structurally coherent narrative.

A common approach for improved summarization is to filter the input (Chen and Bansal, 2018; Pilault et al., 2020; Krishna et al., 2023). As shown in Figure 1, given a stream of flooding event-related documents containing causes alongside impacts, forecasts, and contextual information (e.g., news and social media posts), a pipeline would retrieve cause-relevant content and then pass these documents directly to an LLM for summarization (Route I - left). A more granular approach adds a cause-specific sentence extraction step (Route II - right). The document-level route retains residual non-causal information, while the sentence-level route provides a more controlled set.

The finer-grained filtering may reduce the salience of noise and distractors but may not resolve all task-relevant issues. Prior work from diverse domains suggest that while LLMs achieve high perceived quality in summaries, they can struggle with deeper structural alignment and knowledge abstraction (Zhou et al., 2025; Zeweniuk et al., 2025; Guo et al., 2026). It remains unclear whether cause-specific sentence extraction meaningfully improves summarization beyond the document-level filtering, and whether the resulting summaries are comparable to the structured human-written reference summaries. This distinction is important for evaluating the readiness of LLMs as a deployable tool to aid post-event reporting. If downstream errors are primarily reduced through improved input selection, development should be focused on the cause classification and extraction. However, if consequential differences from human-written summaries remain under controlled inputs, then summary generation itself requires closer task-specific evaluation prior to deployment. Therefore, we ask the following research questions:

1. How does input granularity (cause-specific documents vs. cause-specific sentences) affect the content and task-relevant quality of LLM-generated flood cause summaries?
2. When provided with cause-specific sentences, how do synthesis behaviors of LLMs structurally and semantically diverge with human writers, and are these divergences systemic across different models?

Existing automatic evaluation metrics only provide partial answers: ROUGE (Lin, 2004) measures lexical overlap, while embedding-based metrics such as BERTScore (Zhang et al., 2020) estimate semantic similarity. Although useful indicators for general quality, these metrics cannot directly assess causal direction (`'The flooding caused the building damage'`), amplified speculations and opinion, leaked emotional information, or responsibility attribution. These dimensions are important in disaster communications because summaries can shape how responsibility and accountability are understood.

In this paper, to answer our research questions, we conduct a two-part study of LLM flood cause multi-document summarization. In Part I, we compare the two routes in Figure 1, to examine how input granularity affects the quality of the LLM-generated cause-focused summaries. In Part II, we control for the input condition by using cause-specific sentences and evaluate how LLM-generated summaries differ from human-written summaries using a funnel multi-dimensional human evaluation focused on descriptive precision, task exclusivity, causal framing and responsibility attribution, and structural quality.

We conduct our experiments on six commonly used LLMs and 69 flood events. Our results show that the cause-specific sentence extraction step helps reduce non-causal information for smaller models. However, several divergences from human-written summaries remain across the evaluated models: LLM-generated summaries are longer and less coherent, contain more extreme descriptors, and are less consistent in capturing governmental responsibility.

2 Related Work

Crisis Summarization.
In crisis informatics, multi-document summarization has been a useful tool for reducing large texts into useful and actionable information. Summarization studies have been primarily focused on situational awareness, rescue prioritization, and report generation (Lei et al., 2025; Nguyen and Rudra, 2022; Seeberger and Riedhammer, 2024), utilizing multi-platform sources such as CrisisFacts (McCreadie and Buntain, 2023), and building on LLMs (Pereira et al., 2023; Seeberger and Riedhammer, 2024). However, post-event cause summarization, particularly in multi-platform and

multi-document settings, remains underexplored. Dela Cruz et al.'s (2026b) study initially benchmarked zero-shot flood cause classification, but the cause summarization capabilities remain an open question.

Biases, Input Noise, and Distractors. A challenge in multi-document and long context synthesis is the sensitivity of the LLMs to biases, input noise and distractors (Krishna et al., 2023; Belém et al., 2025; Levy et al., 2025). LLM summarization inherits social and content-selection biases (Anantaprayoon et al., 2024; Gallegos et al., 2024; Steen and Markert, 2024; Zhang et al., 2023) and position bias in summarization (Chhabra et al., 2024; Olabisi and Agrawal, 2024; Ravaut et al., 2024; Wan et al., 2025), which diminish faithfulness as context size increases. Applying RAG, which limits the context window to only relevant facts, reduces or even eliminates the position bias (Cuconasu et al., 2025). Building on these findings, it stands to reason that limiting the context window to controlled cause-specific sentences would yield improved performance, motivating our investigation into input granularity.

Evaluating Synthesis and Causal Framing. Even if input quality is optimized and resolved, inherent LLM synthesis behaviors can still pose problems. LLMs may selectively retain, omit, or re-frame the causal relationships during summarization (Alessa et al., 2025). Sometimes LLMs exhibit higher framing rates than humans in the political news domain (Pastorino and Moosavi, 2026). They can omit potentially controversial information regarding political topics (Noels et al., 2026) or exhibit censorship bias in generation (Ahmed et al., 2025), withhold politics-related information or produce false claims (Urman and Makhortykh, 2025), and reflect ideological stance of their creators (Buyl et al., 2026). This is critical in disaster communications, where news media coverage frames how responsibility is attributed, through narrative and editorial choices, i.e., whether they emphasize natural triggers, human actors, or institutional failures (Bailon et al., 2025; Ewart and McLean, 2015), and where LLMs already struggle to capture temporal and causal links on multiple platforms e.g., news and social media (Zhang and Kou, 2025).

Specific to flood cause extraction, initial benchmarking revealed that LLMs struggle in both cause document classification and cause sentence extraction; they have high recall and low precision by misclassifying general flood-related text as causal (Dela Cruz et al., 2026b). Subsequently, qualitative analysis by Dela Cruz et al. (2026a) demonstrated that false positives produced by LLMs can be clustered into predictable, linguistically coherent categories: impact and aftermath—confusing impacts with causes; background context—misinterpreting general background context of the affected location; forecasts and warnings—stumbling on predictive weather forecasts; subjective and unverified causality—accepting subjective claims and opinion; weak causality—extracting weak causal claims; and unrelated mentions—failing to differentiate natural hazard context use of the word "flood". Appendix B.1 describes these categories in detail. We adopt this exact error taxonomy to answer our first research question.

While LLM summaries can achieve perceived quality in domains like news (Zhang et al., 2023) and medicine (hospital discharge summaries and clinical text summaries) (Van Veen et al., 2024; Challener et al., 2026), they can struggle in structural alignment and knowledge abstraction. Also, deeper structural and synthesis differences continue to persist. For dialog summarization, Zhou et al. (2025) found that LLM summaries are longer and not lexically and structurally aligned to human references. They were nearly five times more likely to contain broad generalizations in science summaries (Peters and Chin-Yee, 2025). They excel at paraphrasing but struggle in reasoning-based and conclusion-oriented summarization (Zeweniuk et al., 2025) and can lead to lower actual reader comprehension (Guo et al., 2026). These mixed findings highlight that comparisons of LLM and human summaries are highly domain- and task-specific.

Necessity for Targeted Human Evaluation. Standard automatic metrics can be blind to the discourse-level requirements of a flood cause summary. ROUGE (Lin, 2004) measures lexical overlap while contextual embedding-based metrics, such as BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021) measure semantic similarity, high scores in these metrics do not automatically reflect instruction-following or accurate responsibility attribution (Zhou et al., 2025; Min et al., 2025) and differ greatly under different human evaluation protocols (Liu et al., 2023; Zhou et al., 2025). This is why there has been strong calls for robust, qualitative and multi-dimensional human evaluation to address these gaps (Lux et al., 2020; Liu et al., 2023; Min et al., 2025). Conse-

quently, the flood cause summarization task necessitates a targeted, multi-dimensional human evaluation to investigate whether synthesis divergences are isolated or are fundamental.

3 Method

3.1 Models

We evaluate six commonly used LLMs: GPT 4o-mini (OpenAI, 2024a), GPT 4o (OpenAI, 2024b), GPT 5-mini (OpenAI, 2025), GPT 5 (OpenAI, 2025), Llama 3.1-8B Instruct (Llama Team, 2024), and Mistral Nemo (Mistral AI Team, 2024) that are accessible to both researchers and practitioners. They were selected to include both closed and open-weight systems, enabling comparison across model families. More details are found in Appendix A.3.

3.2 Data

We run our experiments in a private flood cause dataset comprised of a curated set of documents from four online platforms: News, X (formerly Twitter), YouTube, and Bluesky reporting on 70 flood events that occurred between 2022-2025 as identified and defined by a proprietary retrieval algorithm collected from Dela Cruz et al.'s (2026b) study. Although the dataset cannot be shared, the collection and annotation process are documented in Dela Cruz et al.'s (2026b) paper. The dataset was annotated on two levels of granularity: document-level and sentence-level annotations. For documents identified as containing a flood cause, annotators selected specific sentences that explain the cause of the flood and classify these cause-specific sentences to specific categories. Only the set of cause-specific documents and sentences were used to construct the summaries.

We constructed the structured human-written reference summaries for each flooding event to establish an evaluation baseline. Four trained Communication undergraduate students with English proficiency wrote the summaries based on the requirements of a cause-focused summary. The 70 flooding events were divided equally among students, of which one event was considered invalid due to insufficient information. The summarization writing protocol is in Appendix A.1.

3.3 Prompts

We evaluate the zero-shot performance of the LLMs in the flood cause summarization task. Our prompts were constructed with reference to the flood cause summary annotation protocol (Appendix A.1) and the general guidelines of writing a good quality summary. The full prompts can be found in Tables 7 and 8 in Appendix A.2.

3.4 PART I: Input Granularity of Multi-document Flood Cause Summarization

We evaluate the summaries from the two input conditions using both automatic and human evaluation. First, we evaluate the summaries using standard automatic metrics (1) N-gram overlap using ROUGE-2 and ROUGE-L (Lin, 2004) and (2) semantic similarity using BERTScore (Zhang et al., 2020), by comparing the LLM summaries with the human-written summaries.

To examine whether the residual non-cause content propagates into the LLM-generated summaries and if the cause-specific extraction helps reduce them, we conducted a blinded manual annotation of summary sentences with non-cause content into the false positive error categories of LLMs in the flood cause classification task using Dela Cruz et al.'s (2026a) error taxonomy of flood cause false positives as discussed in Section 2. We count the number of non-cause sentences per error category. As summary sentences can include mixed information, we only count errors for sentences that do not contain any flood cause; annotators were blinded to the input condition and model. Appendix B.1 reports the full description of the error categories and B.3 reports the annotation agreement scores.

3.5 PART II: Task-Specific Evaluation of Flood Cause Summaries

We answer the second research question using a two-stage funnel human evaluation. Our objective with this targeted evaluation is not to fully explore benchmark performance across all models. Rather, we aim to conduct a targeted investigation to identify where systemic divergences with human-written reference summaries still persist. Guided by prior work as discussed in Section 2, we conduct a multi-dimensional approach to ensure holistic evaluation that addresses both the linguistic nuances of causal language and the structural requirements of disaster reporting. Table 1 summarizes the four dimensions, their variables, operationalization, and motivation. Full definitions and annotation agreement scores are provided in Appendix B.2 and B.3.

Stage 1. Blinded Baseline Evaluation

Four trained annotators conducted a blinded de-

Dimension	Variables (count)	Operationalization	Motivation
Descriptive precision	Numerical mentions; extreme descriptors; location, time and date specificity (5)	Counts; Specifity scale (0-3, 3 is most specific); Binary presence (date)	Objective grounding, precise post-event reporting
Task Exclusivity	Casualty and injured; infrastructure damage; emotional reactions (3)	Binary presence of impact-only information	Cause-focused instruction-following
Causal Framing and Responsibility Attribution	Natural, structural, political causes; annotation protocol cause categories; individual, community, governmental responsibility (12)	Binary Presence for each category/frame	Cause frame, explanation, and accountability
Structural Quality	Relevance; coverage; coherence; verbosity (4)	Ratios (no. of sentences/ total sentence count); Likert rating (coherence); total sentence count	Standard human evaluation of summaries

Table 1: Task-specific dimensions for the blinded human evaluation. These dimensions combine the requirements for flood cause-focused summaries and standard summary quality requirements motivated in Section 2; full description can be found in Appendix B.2.

ductive annotation of both human and LLM summaries i.e., authorship was hidden and order of presentation randomized. Each summary was evaluated across more than 20 distinct linguistic and structural variables, operationalized into four task-specific dimensions as shown in Table 1. We compare the baseline LLM and human-written summaries to identify the dimensions on which the model most clearly diverges from the human reference condition. GPT-4o-mini was chosen as the baseline because it showed the strongest automatic metrics alignment with the human reference summaries in Part I and demonstrated significant reductions in residual non-causal content when provided with the idealized, sentence-level inputs. Hence, it provides a useful test case for examining whether synthesis-level divergences remain after controlling input quality.

Stage 2. Targeted Cross-Model Validation

We identify variables on which the baseline model has statistically significant difference from human-written references in Stage 1, then evaluate the remaining models only on this subset. This funnel tests whether the observed divergences recur across the evaluated models rather than being specific to the baseline model.

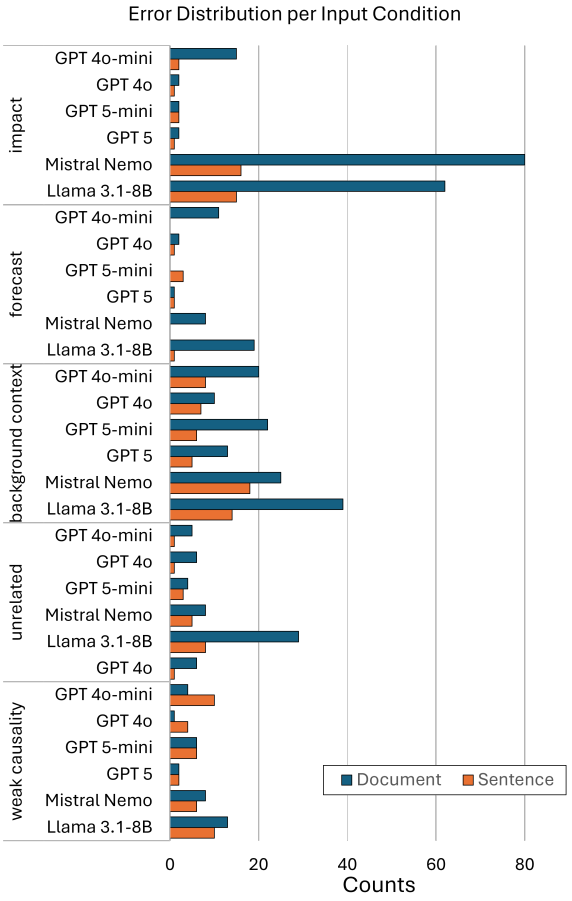


Figure 2: Generation error counts in the summaries for input conditions cause-specific documents vs. cause-specific sentences.

4 Results

4.1 PART I. Input Granularity

Standard automated summary metrics favor the sentence-level input condition. The automated evaluation (ROUGE-2, ROUGE-L and BERTScore) reveal that LLM-generated summaries from the cause-specific sentences are more closely aligned with human-written references than the cause-specific documents as input. Across these metrics and models, the sentence-level input condition consistently achieves higher scores, while the document-level produces significantly longer summaries for all models except for Mistral Nemo.

Measure	Model	Document (M)	Sentence (M)	p-Value
Word Count	GPT 4o-mini	97.2	75.2	<.001
	GPT 5-mini	146.1	98.5	<.001
	GPT 4o	73.7	66	<.05
	GPT 5	109	92.3	<.05
	Llama 3.1-8B Instruct	154.2	119.1	<.05
	Mistral Nemo	75.7	64.8	0.17
ROUGE-2	GPT 4o-mini	**0.12**	**0.17**	<.001
	GPT 5-mini	0.09	0.15	<.001
	GPT 4o	0.14	0.16	<.05
	GPT 5	0.11	0.15	<.001
	Llama 3.1-8B Instruct	0.11	0.17	<.001
	Mistral Nemo	0.08	0.13	<.001
ROUGE-L	GPT 4o-mini	0.22	**0.26**	<.001
	GPT 5-mini	0.18	0.25	<.001
	GPT 4o	**0.24**	0.26	<.05
	GPT 5	0.20	0.24	<.001
	Llama 3.1-8B Instruct	0.18	0.24	<.001
	Mistral Nemo	0.18	0.23	<.001
BERTScore	GPT 4o-mini	0.88	**0.89**	<.001
	GPT 5-mini	0.86	0.88	<.001
	GPT 4o	**0.88**	0.89	<.001
	GPT 5	0.86	0.88	<.001
	Llama 3.1-8B Instruct	0.87	0.88	<.001
	Mistral Nemo	0.85	0.87	<.001

Table 2: Comparison between summary input conditions of cause-specific documents and cause-specific sentences using automatic measures (Word Count, ROUGE-2, ROUGE-L, and BERTScore). The values are measured as Mean. Wilcoxon signed-rank test was applied to the two input conditions.

Model	Impact	Forecast	Background Context	Unrelated Mentions	Weak Causality
GPT 4o-mini	✔	✔	✔	—	—
GPT 4o	—	—	—	—	—
GPT 5-mini	—	—	✔	—	—
GPT 5	—	—	—	—	—
Mistral Nemo	✔	✔	✔	—	—
Llama 3.1 - 8B	✔	✔	✔	✔	—

Table 3: Statistical significance of generation error count differences between input conditions: document-level vs. sentence-level. (✔)indicates a statistically significant difference in errors using the Wilcoxon signed-rank test at $\alpha = 0.05$, while (—) indicates no significance difference.

The full results table can be found in Table 2. Automatic metrics only serves general indicator of overall alignment and confirmation of the presence of more distraction. Figure 2 shows the distribution of error counts for all the error types identified from Dela Cruz et al.'s (2026a) categories.

Human evaluation reveals that input granularity affects smaller models more, while larger models do not benefit from the cause-specific extraction step. Figure 2 and Table 3 show that sentence-level input significantly reduced impact, forecast, and background-context errors for GPT 4o-mini, Mistral Nemo, and Llama 3.1-8B. We observed that the smaller and open-weight models tested (GPT 4o-mini, Mistral Nemo, and Llama 3.1-8B), the document-level condition introduced significantly more residual non-causal content, specifically in the form of impact information, forecasts, and background context. For these models, the cause-specific sentence extraction step helps improve flood cause summarization.

On the other hand, the larger and more advanced models (GPT 4o and GPT 5) demonstrated no statistical significant difference in error counts. However, as observed in Figure 2, this does not mean the larger models suppressed all residual non-causal content, e.g., background context errors still persist. Instead, this indicates that they no longer benefit from the finer-grained cause-specific sentence extraction step. These findings point that there are limitations of LLMs within the synthesis process itself and not the input quality alone.

Dimension	Measure (N=69)	Human	LLM	p-Value
Descriptive Precision	Number of Numerical Mentions (mean)	0.49	1.17	<.001
	Number of Extreme Descriptors (mean)	2.29	4.17	<.001
	Location Specificity (mean scale 0 – 3, 3 is most specific)	2.33	2.38	0.28
	Time Specificity (mean scale 0 – 3, 3 is most specific)	1.14	0.91	0.06
	Date Specificity	49	27	<.001
Task Exclusivity	Casualty and Injured Information	4	4	-
	Infrastructure Damage Reports	14	22	0.12
	Emotional Reactions	14	15	1.00
Causal Framing and Responsibility Attribution	Natural Cause	69	69	-
	Structural Cause	26	25	1.00
	Political Cause	7	0	<.05
	Individual Responsibility	0	0	-
	Community Responsibility	2	1	1.00
	Governmental Responsibility	13	3	<.05
	Rain/Tide/Snow	69	69	-
	Overflowing Water Bodies	15	19	0.50
	Infrastructure Failure	9	6	1.00
	Drainage/Blockage/Maintenance Issues	16	13	0.55
	Long-term Developments	3	2	1.00
	Other Cause	12	6	0.07
Structural Quality	Coherence (mean - 5-point Likert scale)	4.09	3.04	<.001
	Coverage (mean ratio)	0.63	0.40	<.001
	Relevance (mean ratio)	0.86	0.87	0.95
	Verbosity (mean - total sentence count)	3.36	5.42	<.001

Table 4: Evaluation of Human and LLM Summaries (GPT 4o-mini) across the four task-specific dimensions for a cause-focused summary. For the variables in Task Exclusivity and Causal Framing and Attribution dimensions, and Date Specificity, they are counted as total number of summaries with presence of the variable; McNemar tests were conducted on these variables. For the rest of the continuous variables, paired samples t-tests were conducted.

4.2 PART II. Synthesis Analysis and Funnel Evaluation

Controlling for sentence-level input aligns LLMs with humans on relevance and task exclusivity, but does not eliminate synthesis divergences. Table 4 shows the comparison between human and LLM in their summaries of flood causes across the task-specific dimensions. The LLM summaries were provided with the controlled sentence-level conditions. We observed similar performance in the task exclusivity dimension, where both human and LLM leak casualty and injured information, infrastructure damage reports, and emotional reactions in the summaries; there were no statistically significant differences observed. Our results show that finer-grained extraction alone does not fully resolve the core challenges of the task.

The largest differences emerged in the descriptive precision dimension. The LLM averaged more numerical mentions compared to humans (1.17 vs. 0.49, $p < .001$) and nearly doubled the number of extreme descriptors (4.17 vs 2.29, $p < .001$) in the summaries. We also observed the LLM summaries contain less specific dates (27 vs. 49 date-specific mentions, $p < .001$).

Structural quality differences were observed in coherence, coverage and verbosity. The LLM summaries were significantly longer than the human summaries (M = 5.42 vs. 3.36 sentences, $p < .001$). The longer summary length did not translate into more information coverage, in fact, the LLM summaries had a significantly lower coverage ratio (M = 0.40 vs. 0.64, $p < .001$). Lastly, the LLM summaries were significantly less coherent (M = 3.04 vs. 4.09, $p < .001$) than human summaries.

Both human and LLM summaries consistently identified the natural or environmental causes with recognizing the natural cause framing and the Rain/Tide/Snow variable appearing in both groups. Key differences emerged in the higher-level institutional factors, where the LLM failed to capture political causes (0 vs. 7, $p < .05$) and failed to attribute governmental responsibility (3 vs. 13, $p < 0.05$). Table 5 shows an illustrative example.

Cross-model funnel evaluation reveals some synthesis divergences to be systemic and not just model-specific. Table 6 shows the results of the targeted cross-model validation. We observed that every LLM produced significantly more extreme descriptors (EXTR) than the human summary. Numeric mentions (NUM) had significant over-representation in GPT 4o-mini, GPT 5-mini and GPT 5. Date specificity (DATE) continues to be a challenge with only the latest models, GPT

Human Summary	LLM Summary
On Friday 7 March, a powerful storm and heavy rains brought catastrophic flooding to Bahía Blanca, Argentina. The intense rain started at around 4 am and lasted about 7–8 hours. 400 mm of rain fell, which is equivalent to a year's worth. The World Weather Attribution analysis revealed that extreme regional heat clashed with a cold front, resulting in heavy rainfall partly fueled by climate change, which caused the flooding. **Also, the flooding was caused by poor drainage infrastructure, and responsibility for this falls to the provincial government.**	Torrential rains caused severe flooding in Bahia Blanca, Argentina, on March 7, 2025. More than 400 mm of rain fell in just a few hours, leading to catastrophic flooding. A year's worth of rain fell in a matter of hours, inundating entire neighborhoods. Extreme heat and humidity preceding the floods contributed to the intense rainfall. A collision between warm and cold air masses triggered the massive rainfall.

Table 5: Example comparison of flood cause summaries for a flooding event in Buenos Aires in March 2025 from a human summarizer and an LLM (GPT 4o-mini). Sentence in RED indicates a type of causal frame and responsibility attribution in the human summary not captured by the LLM.

5-mini and GPT 5 matching the human summaries.

For most LLMs, catching political causes (POL) is on par with human summaries except for GPT 4o-mini and Llama 3.1. However, there is a consistent failure in the attribution of governmental responsibility (GOV) for LLMs. Quite alarmingly, newer models, specifically, GPT 5-mini and GPT 5 generated no mention of implicating governments. **Summary structural quality differ between LLMs and humans in the cause summarization task.** Across the board, LLM generated longer summaries (VERB) but their coherence (COH) scores were significantly lower than the human summaries. Newer models, GPT 5-mini have surpassed the human coverage ratio (COV) while GPT 4o-mini and Llama 3.1 had lower coverage ratios.

5 Discussion and Conclusion

Our findings suggest that the cause-specific sentence-level extraction step improves the LLM performance in the flood cause summarization by reducing possible error categories for smaller and open-weight models, while reductions are not significant for the larger ones. We note that the automatic metrics results should be interpreted cautiously as the human summaries used the cause-specific sentences set, which makes them more naturally aligned with the sentence-level input condition, hence, human evaluation is needed. The human evaluation showed that the full documents, although already cause-relevant, contained highly salient distractors for LLMs, mostly impact information and disaster background context which are significantly reduced by the idealized sentence extraction step. A cause-focused post-event report that leaks all the impact information or background context data would require more human review by disaster managers.

Despite the highlighted synthesis divergences, it is important to acknowledge the significant baseline capabilities of the LLMs in the flood cause summarization task. Among the more than 20 linguistic and structural variables, approximately two-thirds of the variables converged with human writers. First, LLMs aligned closely with humans on relevance, meaning they provide actual information need for the task which are flood causes. This is further strengthened by the convergence in task exclusivity, highlighting the inherent difficulty of omitting impact information and background context when summarizing for the causes. Furthermore, both humans and LLMs consistently identified natural and environmental causes.

However, removing residual noise does not automatically eliminate all divergences. For the remaining divergent variables, the LLMs functioned as selective amplifiers. We observed that they were systematically longer and over-generate numerical mentions and extreme descriptors compared to humans. These findings align with the study by Alessa et al. (2025), who found that LLMs often alter text framing to more negative. If deployed for post-event reporting, LLMs can post a risk in disaster communications by making causes appear to be more catastrophic or extreme than they are.

Conversely, the evaluated LLMs were found to under-represent higher-level governmental responsibility, with some found to fail to capture political causes. This finding serves as possible evidence for the "soft censorship" phenomenon (i.e., selective omission or downplaying of key elements) identified by Noels et al. (2026) or avoidance of

Measure	Human (Reference)	GPT 4o-mini	GPT 4o	GPT 5-mini	GPT 5	Llama 3.1-8B	Mistral Nemo
NUM	0.49	1.17	0.48	0.93	0.88	0.64	0.38
EXTR	2.29	4.17	4.96	6.61	5.42	7.45	4.25
DATE	49	27	16	50	54	28	16
POL	7	0	6	5	5	7	2
GOV	13	3	2	0	0	1	1
COH	4.09	3.04	3.46	3.58	3.49	2.69	2.73
COV	0.63	0.40	0.72	0.79	0.86	0.50	0.65
VERB	3.36	5.42	4.72	4.83	4.71	6.20	4.52

Table 6: Funnel evaluation of LLM summaries across initial divergent dimensions. Values in RED and BLUE indicate significant difference from the human-written summaries (color difference indicate different direction).

politics-related information identified by Urman and Makhortykh (2025), demonstrating how LLMs can omit politically sensitive information during synthesis. In a real-world deployment scenario, these LLM-generated post-event reports serve as complementary or supplementary to the official post-event reports, which are used for legal accountability, insurance settlements, and long-term mitigation planning. If a deployed LLM sanitizes reports by omitting mentions of government responsibility, such as `'drainage mismanagement'` or `'ignored dam maintenance'`, the harm is not merely on the data inaccuracies, but it directly impacts citizens. This create artificial narratives that floods are purely unavoidable natural disasters rather than preventable governance failures.

6 Future Work

Our study investigated whether LLM synthesis limitations in the multi-document flood cause summarization task persist when models receive prefiltered, cause-specific inputs. Future research is needed to evaluate the readiness of using LLMs as deployable tools for disaster post event reporting.

Within crisis informatics post-event reporting, future work should test whether the finer-grained sentence extraction step similarly reduces residual non-causal noise for other disaster types and languages. It should also examine whether the structural and semantic divergences observed here persist when summarizing causes of events such as wildfires or landslides.

Although this study focused on flooding events, the observed variable divergences with human-written reference summaries may extend to other forms of query-based multi-document summarization. Our multi-dimensional framework can serve as an initial guide for the domain and task-relevant evaluation but should be validated with actual DRM practitioners and domain experts. Future work can examine whether our identified task-specific variables transfer to other accountability or responsibility-sensitive domains, such as legal, medical, or public policy summarization. Particularly, whether there are tendencies to over-generate extreme descriptors the failure in capturing political causes and governmental attribution.

Finally, our zero-shot design was intended to establish the default behavior of off-the-shelf LLMs in this task. Future research should investigate practical methods for mitigating the critical divergences identified here. Because this study is intended to support practitioners without extensive AI expertise, prompt-tuning is considered the relevant first direction. The evaluation framework introduced in this study can then be used to assess whether revised prompts improve not only general summary quality, but also descriptive precision, task exclusivity, causal framing, responsibility attribution, and structural quality.

Limitations

This paper has limitations. Our dataset is a curated private dataset in English collected from multiple platforms: News, X (formerly Twitter), YouTube, and Bluesky using the data collection and annotation from (Dela Cruz et al., 2026b). This dataset was collected during 70 specific flooding events between 2022 and 2025 as identified and defined by a proprietary retrieval algorithm. Although the dataset cannot be shared, the collection and annotation process are documented in Dela Cruz et al.'s (2026b) paper.

The scope of the use case is very specific to flooding event cause summarization. We did not conduct analyses on other disaster events. Our multi-dimensional approach to domain and task-specific human evaluation was scoped for our specific use case and not the full requirements of a flood post event report. We acknowledge that this

scope is limited and encourage other researchers evaluate on other disaster events and other report types. In addition, we only evaluated zero-shot LLM summarization on six off-the-shelf LLMs. Prompt-tuning and fine-tuning the models to our specific task should be explored and investigated.

The human reference summaries were written by four Communications undergraduate students with English proficiency, who were trained and guided by someone with crisis informatics expertise and not written by actual DRM domain experts. The subsequent summary annotations for human evaluation were conducted by four Communication undergraduate students (note: summary writer was not summary evaluator), and two authors (both are fluent in English, one graduate student with computational linguistics, NLP, and crisis informatics expertise and another graduate student specialized in computer science and AI). We refer you to Appendix B.3 for the inter-annotator agreement scores. The annotation scope was very focused due to the sensitive nature of the dataset. Although already filtered to documents containing causes to the flooding event, the dataset still contains distressing crisis content of impacts of flooding event such as reports of deaths, injuries, missing persons, rescue needs, infrastructure damage, etc.

Ethical Considerations

This study highlights concern about the real-world deployment of LLMs in the disaster risk management (DRM) domain. Our work aspires to support DRM practitioners by demonstrating and clarifying the capabilities of using LLMs in zero-shot settings for flood cause summarization. Evaluated LLMs fed with the idealized input conditions exhibit significant alignment across over two-thirds of the variables investigated that include relevance and task exclusivity variables (i.e. following the task). However, LLMs in this setting do not achieve the required precision and objectivity for unsupervised deployment. Specifically, we observed systemic over-generation of extreme descriptors and numerical information. They also under-represent governmental responsibility and some also under-represent political causes. Therefore, we caution against misinterpreting the readiness of the models investigated for automated flood cause summarization or integration in post-event reporting.

However, we also caution against premature abandonment of the use of LLMs completely for the solutions we study. The LLMs investigated already has considerable alignment in multiple variables, especially relevance. As the divergence with human summaries are measurable, this means they are amenable for targeted improvement with domain experts. They may already add value in the human-in-the-loop settings where our DRM practitioners maintain full interpretive control. Therefore, we strongly recommend only supervised use, using LLMs as complementary and supplementary tools, rather than autonomous report generators.

We recognize that the annotation work for our task carries emotional burden. Although our data was already pre-filtered to be relevant to a flooding event and containing underlying flood causes, the source documents and sentences contained possible highly distressing information regarding the context and impacts of these real-world events. Some documents included reports of injuries, casualties, missing persons, and widespread infrastructure damage. Asking our annotators with reading and summarizing these content carries inherent psychological risks. Annotators were informed and warned beforehand about the sensitive nature of the task prior to participation. They were encouraged to take breaks or opt out if the task became overwhelming.

Acknowledgments

This publication is part of the project 'Indeep: Interpreting Deep Learning Models for Text and Sound' with project number NWA.1292.19.399, which is partly financed by the Dutch Research Council (NWO). We would like to thank FloodTags for their collaboration—for providing the case studies and dataset. We would like to thank Radboud University International Business Communication and Global Communication and Diversity students who were our annotators: Svetlana Monheimus, Josine Meintein, Nikhil Prashar, and Gregory Rizopolous.

References

Mohamed Ahmed, Jeffrey Knockel, and Rachel Greenstadt. 2025. An analysis of Chinese censorship bias in LLMs. *Proceedings in Privacy Enhancing Technologies Symposium*, 2025:112–129.

Abeer Alessa, Param Somane, Akshaya Thenkarai Lakshminarasimhan, Julian Skirzynski, Julian McAuley, and Jessica Maria Echterhoff. 2025. Quantifying cognitive bias induction in LLM-generated content. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th*

Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2890–2910, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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

Hannah Bailon, Kees Boersma, Claudia Orellana-Rodriguez, and Marc Van Den Homberg. 2025. Framing of disaster impact in online news media: a case study from Malawi on flood risk management. *Frontiers in Communication*, Volume 10 - 2025.

Catarina G Belém, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. From single to multi: How LLMs hallucinate in multi-document summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5291–5324, Albuquerque, New Mexico. Association for Computational Linguistics.

M. Buyl, A. Rogiers, S. Noels, G. Bied, I. Dominguez-Catena, E. Heiter, I. Johary, A.-C. Mara, R. Romero, J. Lijffijt, and T. De Bie. 2026. Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2:7.

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

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

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

Florin Cuconasu, Simone Filice, Guy Horowitz, Yoelle Maarek, and Fabrizio Silvestri. 2025. Do RAG systems really suffer from positional bias? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28022–28036, Suzhou, China. Association for Computational Linguistics.

Jane Arleth Dela Cruz, Iris Hendrickx, and Martha Larson. 2026a. From errors to precision: A taxonomy-guided approach for LLM flood cause extraction. In *Proceedings of the Eleventh Workshop on Noisy and User-generated Text*. Association for Computational Linguistics. To appear.

Jane Arleth Dela Cruz, Aleksei Simonov, Iris Hendrickx, and Jurjen Wagemaker. 2026b. Extracting the 'why': Flood cause annotation and LLM benchmarking from multi-platform crisis data. *Proceedings of the 23rd International Conference on Information Systems for Crisis Response and Management*.

Jacqui Ewart and Hamish McLean. 2015. Ducking for cover in the 'blame game': news framing of the findings of two reports into the 2010–11 queensland floods. *Disasters*, 39(1):166–184.

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

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

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Yue Guo, Jae Ho Sohn, Gondy Leroy, and Trevor Cohen. 2026. Are llm-generated plain language summaries truly understandable? a large-scale crowdsourced evaluation. *Journal of Biomedical Informatics*, 179:105038.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics.

Muhammad Imran, Abdul Wahab Ziaullah, Kai Chen, and Ferda Ofli. 2025. Evaluating robustness of LLMs on crisis-related microblogs across events, information types, and linguistic features. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 5117–5126, New York, NY, USA. Association for Computing Machinery.

Kundan Krishna, Yao Zhao, Jie Ren, Balaji Lakshminarayanan, Jiaming Luo, Mohammad Saleh, and Peter Liu. 2023. Improving the robustness of summarization models by detecting and removing input noise. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1324–1336, Singapore. Association for Computational Linguistics.

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

Shahar Levy, Nir Mazor, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. 2025. More documents, same length: Isolating the challenge of multiple documents in RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19539–19547, Suzhou, China. Association for Computational Linguistics.

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

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

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

Klaus-Michael Lux, Maya Sappelli, and Martha Larson. 2020. Truth or error? towards systematic analysis of factual errors in abstractive summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 1–10, Online. Association for Computational Linguistics.

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

Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multi-dimensional evaluation of LLM summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450, Vienna, Austria. Association for Computational Linguistics.

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

Thi Huyen Nguyen and Koustav Rudra. 2022. Rationale aware contrastive learning based approach to classify and summarize crisis-related microblogs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1552–1562, New York, NY, USA. Association for Computing Machinery.

Sander Noels, Guillaume Bied, Maarten Buyl, Alexander Rogiers, Yousra Fettach, Jefrey Lijffijt, and Tijl De Bie. 2026. What large language models do not talk about: An empirical study of moderation and censorship practices. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 265–281, Cham. Springer Nature Switzerland.

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

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

OpenAI. 2024b. Gpt-4o system card.

OpenAI. 2025. Gpt-5 system card.

Valeria Pastorino and Nafise Sadat Moosavi. 2026. Frame in, frame out: Measuring framing bias in LLM-generated news summaries. In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 378–384, San Diego, California, United States. Association for Computational Linguistics.

J. Pereira, R. Fidalgo, R. Lotufo, and R. Nogueira. 2023. Crisis event social media summarization with GPT-3 and neural reranking. In *Proceedings of the 20th International ISCRAM Conference*, pages 371–384, Omaha, Nebraska, USA.

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

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

pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Seeberger and Korbinian Riedhammer. 2024. Crisis2Sum: An exploratory study on disaster summarization from multiple streams. *Proceedings of the International ISCRAM Conference*, 21.

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

Aleksandra Urman and Mykola Makhortykh. 2025. The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in chatgpt, google bard (gemini), and bing chat. *Telematics and Informatics*, 96:102211.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8791–8810, Albuquerque, New Mexico. Association for Computational Linguistics.

Fengyi Xu, Jun Ma, Nan Li, and Jack C.P. Cheng. 2025. Large language model applications in disaster management: An interdisciplinary review. *International Journal of Disaster Risk Reduction*, 127:105642.

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

Nathan Zeweniuk, Ori Ernst, and Jackie CK Cheung. 2025. Beyond paraphrasing: Analyzing summarization abstractiveness and reasoning. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 48–58, Hybrid. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via ChatGPT for faithful summary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3270–3278, Singapore. Association for Computational Linguistics.

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

Wei Zhang and Yang Kou. 2025. Temporal and causal-aware summarization of legal public opinion with large language models. In *2025 IEEE 8th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 777–785.

Yongxin Zhou, Fabien Ringeval, and François Portet. 2025. Can GPT models follow human summarization guidelines? a study for targeted communication goals. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 249–273, Hanoi, Vietnam. Association for Computational Linguistics.

A Additional Method Details

A.1 Human Flood Cause Summary Annotation Protocol

Writing the flood causes summary

Write a short summary that includes all the causes mentioned across the documents for a flooding event.

You are provided with sentences that are classified to contain flood causes. Read and understand the sentences. Select the most informative sentences about the flood cause that you would include in your summary. Mark these sentences on the table. Use your selected informative sentences to write your short summary.

Qualities of a Good Summary:

- **Consistency:** Use only information found in selected source documents. Please do not add new information, speculate, or add personal opinion.
  - Example: If documents or sentences mention "drainage blockage" but do not specify a reason (trash accumulation, etc.), do not add one.
- **Relevance:** Focus exclusively on the flood causes. Do not include the impacts such as injuries, evacuations, damages etc., or future plans or responses.
- **Coverage:** Include all causes mentioned across selected documents. Avoid unnecessary repetition. Aim for one sentence per unique flood cause when applicable. Make sure all your marked sentences are covered.

- Report the cause of the flooding event: e.g., 1) natural causes (heavy rainfall, extreme weather conditions, snowmelt, climate-related factors) or 2) structural/infrastructural causes (dam failures, inadequate drainage systems, infrastructure weaknesses) or 3) political/governmental causes (policy decisions, regulatory failures, lack of preparedness, governmental mismanagement). There could be more than one.
- Add critical information like location (go as specific as possible), date (if specific dates are not mentioned, it can be last week, last month too), and time (14:00, early morning, late evening, etc.) of the event.

- **Coherence:** A good summary is a clear, readable, logical narrative. Ensure your final summary reads as a brief report explaining what caused the flooding in past tense.

Final Check

- Proofread your summary for clarity, grammar, spelling, and sentence flow.
- Confirm that you covered all causes, included no non-cause details, and added no information not present in the documents. Review from your chosen important sentences.

A.2 Prompts

The prompts used in our paper are found in Tables 7 and 8.

A.3 Models

Table 9 contains the information about the versions of the six LLMs we evaluated and analyzed. All six models, except for GPT 5-mini and GPT 5, were set at temperature 0.0 to be deterministic. Both GPT 5-mini and GPT 5 models were set to reasoning effort "low."

System	You are a concise assistant. Use ONLY the given data. No explanations or thought process. Just the final answer.
User	You are provided with documents that are classified as containing information about the cause of the flood. Task: Write a short bullet-list summary that includes all the causes explicitly mentioned across the documents for the flooding event. Steps: 1. Select the most informative documents about the flood cause. 2. Use the selected the documents to write your short summary. 3. Qualities of a good summary: a. Consistency - use only information in the selected documents. Do not add new information, speculate, or add personal opinion. b. Relevance - focus exclusively on flood causes. Do not include impacts such as injuries, evacuations, damages, etc., or future plans or responses c. Coverage - include all causes mentioned in the selected documents. Avoid unnecessary repetition. Aim for one sentence per unique flood cause. Report the causes and critical information like location, date, and time. d. Coherence - make your summary clear, readable, and logical. It must read as a brief report explaining what caused the flooding in past tense. 4. Output your summary as a bullet-list. Documents: **{DOCUMENT LIST}**

Table 7: Summary Prompt for Cause-Specific Documents as Input. **{DOCUMENT LIST}** is the list of cause-specific sentences for a flooding event. The list is sorted by timestamp.

B Task-Specific Evaluation

B.1 Flood Cause Classification Error Categories

In their findings, LLMs misinterpret **Impact and Aftermath** information, such as "`Residents 'evacuated' from rural homes near Chester due to flooding`", by treating impact descriptions (outcomes), like damage or evacuation as causes due to presence of causality but invert the cause-and-effect relationship. Similarly, LLMs mistake general **Background Context** for specific flood event triggers, overgeneralizing location descriptions like, "`Hail storm and a record amount of precipitation and lightning in Valencia city`" even though no flooding event is described. They also stumble on weather **Forecasts and Warnings**, where they falsely interpret predictive language like "`Heavy rainfall today may`

System	You are a concise assistant. Use ONLY the given data. No explanations or thought process. Just the final answer.
User	You are provided with sentences that are classified as containing information about the cause of the flood. Task: Write a short bullet-list summary that includes all the causes explicitly mentioned across the sentences for the flooding event. Steps: 1. Select the most informative sentences about the flood cause. 2. Use the selected the sentences to write your short summary. 3. Qualities of a good summary: a. Consistency - use only information in the selected documents. Do not add new information, speculate, or add personal opinion. b. Relevance - focus exclusively on flood causes. Do not include impacts such as injuries, evacuations, damages, etc., or future plans or responses c. Coverage - include all causes mentioned in the selected documents. Avoid unnecessary repetition. Aim for one sentence per unique flood cause. Report the causes and critical information like location, date, and time. d. Coherence - make your summary clear, readable, and logical. It must read as a brief report explaining what caused the flooding in past tense. 4. Output your summary as a bullet-list. Sentences: **{SENTENCE LIST}**

Table 8: Summary Prompt for Cause-Specific Sentences as Input. **{SENTENCE LIST}** is the list of cause-specific sentences for a flooding event. The list is sorted by document and timestamp.

cause some flash flooding across Scotland" as factual reporting post-event.

In addition to contextual and directional errors, LLMs rely on explicit lexical cues and have broad semantic scope. They accept **Subjective and Unverified** causality, e.g., "Well, at least California has destroyed flood control dams so more natural water flow fills bays so some freaking snail survives" as valid because they fail to differentiate objective reporting from speculation, opinion and sarcasm. Lastly, LLMs can be fooled by superficial keyword matches we call **Unrelated Mentions** and fail to differentiate natural hazard context use of flood from metaphorical or non-hydrological use of the it. For example, an LLM fails in the text: "Eight people die in the London Beer Flood, an accident at Meux & Co's Horse Shoe Brewery" to restrict the semantic scope to hydrological events.

B.2 Multi-dimension Evaluation Variables

- **Descriptive Precision. (5 variables)** LLM-generated summaries can suffer from hallucination and overgeneralizations (Zhou et al., 2025; Belém et al., 2025; Peters and Chin-Yee, 2025). In DRM, an objectively grounded report is critical; therefore, we measure *negativity bias* as counting the use of extreme descriptors and numerical information, together and *specificity of spatiotemporal information* as the exactness of the location and time rated on a 0-3 scale, alongside a binary presence for date information.
  - Numerical Mentions - any explicit numerical expression that refers to causal or contributing factors of the flood event. Numbers referring to consequences must not be coded. Each distinct numerical reference related to flood causes is counted as one instance.
  - Extreme Descriptors - words or phrases that go beyond neutral factual reporting and add evaluative, emotional, or catastrophizing emphasis to the description of the flood. Only explicit linguistic expressions in the summary should be coded; do not infer emotional tone.
  - Location Specificity - measure level of geographic detail preserved in the summary. (0 - no information available, 1

Model	Type	Context Window (tokens)	Access (OpenAI/Huggingface)
GPT 4o-mini	closed	128K	gpt-4o-mini-2024-07-18
GPT 4o	closed	128K	gpt-4o-2024-08-06
GPT 5-mini	closed	400K	gpt-5-mini-2025-08-07
GPT 5	closed	400K	gpt-5-2025-08-07
Llama 3.1 - 8B Instruct	open	128K	meta-llama/Llama-3.1-8B-Instruct
Mistral NeMo	open	128K	mistralai/Mistral-NeMo-Base-2407

Table 9: Information of evaluated and analyzed LLMs

- vague (country level), 2- moderately specific (city, state, province, region), 3 - highly specific (street name, address, building names, landmarks)
  - Temporal Specificity - measure specificity of time related information. (0 - no information available, 1 - recently, this month, this week), 2 - moderately specific (early morning, evening), 3 - highly specific (exact time is mentioned)

- **Task Exclusivity. (3 variables)**. A query-based domain-specific summary must follow strict Instructions. To verify that the strict semantic boundary between the task-requirement is maintained between causes and impacts, we track the *presence of impact information* (e.g., casualties, infrastructure damage, and emotional reactions) in the summaries.
  - Casualty/Injured Information - refers to individuals who are killed, injured, or otherwise physically affected by an incident or disaster, often requiring medical attention or emergency response. Code as (1-Present). Vague statements of "people were affected" should not be included.
  - Infrastructure Damage Reports - refer to documented occurrences of damage caused by an incident, including losses to property and crops and the costs of repairing damaged infrastructure. Code 1 if present, 0 if no material destruction is explicitly described
  - Emotional Reactions - reactions that are a result of certain mental states that include commonly recognized categories such as happiness, sadness, fear, anger, disgust, and surprise. Code 1 if at least one explicit emotional expression is present.

  This was based on the exclusion list from the original dataset's annotation protocol (Dela Cruz et al., 2026b).

- **Causal Framing and Responsibility Attribution. (12 variables)** LLMs are known to selectively retain, omit, or reframe texts (Alessa et al., 2025; Pastorino and Moosavi, 2026; Noels et al., 2026; Urman and Makhortykh, 2025). We evaluate causal framing by annotating the presence of natural causes, structural causes and/or political causes. In addition, we want to know which categories of causes based on the cause categories from (Dela Cruz et al., 2026b) are easy or considered difficult to capture by tracking the presence of these causes in the summaries. For responsibility attribution, we annotate where the accountability is directed at whether it is directed to individual, community, or governmental responsibility reflecting disaster communications (Bailon et al., 2025).
  - Natural Causes - weather or environmental causes are explicitly mention. Code 1 if present; this includes rainfall, storms, monsoons.
  - Structural/Infrastructural - explanations of failures or limits of infrastructure, these physical systems contribute to the flooding. Code 1 if present.
  - Political/Governmental Causes - explanation of governance or policy failure linked to political or institutional decisions. Code 1 if present.
  - Individual Responsibility - individuals explicitly blamed for the event. Code 1 if present.
  - Community Responsibility - collective actors or groups or societies are responsible for the flood. Code 1 if present.
  - Government Responsibility - government or institutions are blamed
  - For the six variables: Rain/Tide/Snow, Overflowing Water Bodies, Infrastructure Failure,

Drainage/Blockage/Maintenance Issues, Long-term Developments, and Other Cause we refer to (Dela Cruz et al., 2026b)'s paper. Code 1 if present

- **Structural Quality. (4 variables)** To assess the overall organization quality of the summaries, we measure *relevance*, *coherence*, and *coverage* which are considered the default variables for human evaluation for summarization (Gehrmann et al., 2018; Fabbri et al., 2021). We also evaluate *verbosity* to assess the differences between LLM and humans using total sentence count. Coherence is measured by a Likert-scale rating for clarity, readability and logical flow. For coverage and relevance, these were counts of unique covered causes and relevant sentences in the summary.

B.3 Inter-Annotator Agreement Scores

We report the inter-annotator agreement scores in Tables 10 and 11. Table 10 show the error category annotation agreement scores. The measures were computed on 84 summaries (2 input conditions X 6 models X 7 flooding events). The agreement scores from Table 11 were computed on 28 summaries (14 flooding events X 2 summary types (LLM and Human)).

Error Category	Reliability (ICC)	Interpretation
impact information	0.65	Moderate
forecast and warnings	0.66	Moderate
background context	0.53	Moderate
unrelated mentions	0.64	Moderate
weak causality	0.59	Moderate
subjective and opinions	NA	Almost Perfect

Table 10: Inter-annotator agreement scores for the error categories. Agreement scores were measured using Intraclass Correlation Coefficient (ICC).

Dimension	Variable	Method	Reliability Score	Interpretation
Descriptive Precision	Numerical Mentions	ICC	0.89	Good
	Extreme Descriptors	ICC	0.71	Moderate
	Location Specificity	Cohen's κ	0.63	Substantial
	Time Specificity	Cohen's κ	0.51	Moderate
	Date Specificity	Cohen's κ	0.71	Substantial
Task Exclusivity	Casualty and Injured Information	Cohen's κ	0.47	Moderate
	Infrastructure Damage Reports	Cohen's κ	0.83	Almost perfect
	Emotional Reactions	Cohen's κ	0.52	Moderate
Causal Framing and Responsibility Attribution	Natural Cause	Cohen's κ	–	Almost Perfect
	Structural Cause	Cohen's κ	0.65	Substantial
	Political Cause	Cohen's κ	1.00	Almost Perfect
	Individual Responsibility	Cohen's κ	–	Almost Perfect
	Community Responsibility	Cohen's κ	0.65	Substantial
	Governmental Responsibility	Cohen's κ	0.84	Almost Perfect
	Rain/Tide/Snow	Cohen's κ	–	Almost Perfect
	Overflowing Water Bodies	Cohen's κ	0.62	Substantial
	Infrastructure Failure	Cohen's κ	0.65	Substantial
	Drainage/Blockage/Maintenance Issues	Cohen's κ	0.79	Substantial
	Long-term Developments	Cohen's κ	1.00	Almost Perfect
	Other Cause	Cohen's κ	0.55	Moderate
Structural Quality	Coherence	ICC	0.63	Moderate
	Coverage	ICC	0.73	Moderate
	Relevance	ICC	0.99	Relevance
	Verbosity	ICC	1.00	Excellent

Table 11: Inter-Annotator Agreement Scores across the four task-specific dimensions.