

Seeing the Storm: Leveraging Multimodal LLMs for Disaster Social Media Video Filtering

Holden Clark

Brigham Young University
holdio@byu.edu

Amanda Lee Hughes

Brigham Young University
amanda_hughes@byu.edu

ABSTRACT

Emergency management relies on the rapid triage of information to respond appropriately to disaster events. Social media platforms can provide emergency managers with ground-level insights, and videos, in particular, offer an immersive medium for understanding public responses and on-the-ground conditions. However, the overwhelming volume of irrelevant or redundant videos complicates their use for emergency response. This paper investigates the use of multimodal large language models (MLLMs)—specifically the Gemini 1.5 flash model—to automate the identification of relevant videos shared on X (formerly Twitter) during hurricanes. We develop and evaluate a framework to test the accuracy of different prompting styles and question strategies. By identifying the most effective prompting techniques, this study lays the groundwork for a systematic approach to filtering social media videos, enabling emergency managers to focus on the most pertinent content and make timely, informed decisions.

KEYWORDS

Crisis Informatics, Emergency Management, Multimodal Large Language Models, Social Media, Video Filtering

INTRODUCTION

Social media have become an important resource for information exchange and decision-making during natural disasters (Palen & Hughes, 2018). People affected by disasters use platforms like X (formerly Twitter), Facebook, and others to assess risks, share updates, and make safety-related decisions (Freeman et al., 2021; Hiltz et al., 2020; Reuter et al., 2018). These platforms can also serve as a tool to communicate the threat level of nearby disasters to others (Freeman et al., 2021; Sutton et al., 2008). The content shared on social media comes in a variety of formats, including text, photos, and video.

Emergency Managers increasingly use social media to support response efforts. They use this user-generated content to identify at-risk populations, assess resource needs, and enhance situational awareness (Palen & Hughes, 2018; Zade et al., 2018). Official emergency management guidelines encourage using real-time, user-generated content to improve response efforts (Purohit et al., 2025). However, the sheer volume and unstructured nature of social media content presents challenges. Emergency managers must sift through vast streams of content to extract relevant information—a process that is both labor-intensive and time-sensitive.

Recent advancements in artificial intelligence (AI), particularly multimodal large language models (MLLMs), offer promising solutions to these challenges. Unlike traditional AI models, MLLMs can process multiple modalities, including text, images, and audio. This capability is particularly valuable for filtering and prioritizing social media videos, which often provide rich, eyewitness accounts that text or images alone cannot fully capture. Videos can document unfolding events, infrastructure damage, and the needs of the disaster-affected public (Schlein et al., 2024). However, despite their potential value, videos remain underutilized due to the manual effort required to review them

and the large presence of irrelevant content, such as memes, generic weather updates, or unrelated footage. These issues make it difficult to efficiently identify videos that may contain actionable information.

This research investigates how MLLMs can help automate the identification and prioritization of relevant social media videos for emergency management. Our study addresses the following research questions:

1. How can MLLMs be used to identify relevant publicly-generated videos for emergency management?
2. How do different prompting styles influence the quality and relevance of MLLM responses for emergency management?

By seeking effective prompting strategies for MLLM-based video analysis, this work aims to streamline the identification of social media videos that could be useful for emergency managers. Automating this process has the potential to enhance situational awareness, improve decision-making workflows, and ultimately increase the speed and efficiency of emergency response efforts.

BACKGROUND

Historically, much of the research on social media in disaster contexts has focused on text-based content (Imran et al., 2015; Reuter et al., 2018). Tweets, hashtags, and status updates are easier to analyze at scale, and studies have demonstrated how such data can be mined to enhance situational awareness (Vieweg et al., 2010; Zade et al., 2018), detect misinformation (Barman & Colan, 2023; Starbird et al., 2014), and inform decision-making processes (Herrera, 2021; Hiltz et al., 2020; Samadi et al., 2024). However, the reliance on text-based data limits the scope of situational awareness, as crucial visual evidence—such as damage assessment—remains largely unaccounted for.

Beyond text, researchers have also explored how social media images can aid disaster response. Imran et al. (2020) used AI-driven image labeling and extraction techniques to classify photos by identifying infrastructure damage or mapping flooded areas. These studies rely heavily on supervised AI techniques that require large amounts of labeled data to train models. While these approaches improve automated classification, their dependence on labeled datasets presents challenges for real-time disaster response, where new and evolving conditions often lack predefined labels.

Building on these efforts, recent research has explored the potential of AI to aggregate and summarize social media data for emergency response. Colverd et al. (2023) used LLMs to automatically generate reports about floods, demonstrating their potential to synthesize large amounts of information into actionable summaries for emergency responders. Taking this further, (Otal et al., 2024) proposed an LLM-enhanced application that integrates social media data with official instructions from emergency managers to provide real-time guidance during disaster events. They found that LLM-generated responses were often impacted by misinformation and irrelevant information found on social media. This findings suggests that filtering out unreliable data could improve response accuracy.

Despite recent advances, LLMs still face substantial challenges in disaster management, especially regarding information accuracy and event identification. Rawat (2024) proposed a benchmark to evaluate LLMs' disaster management. The models performed decently, but not well enough for practical use. Prompting methods improved accuracy, but consistency was lacking. Similarly, Fang (2024) tested prompt engineering techniques to optimize LLM performance across various disaster scenarios. Their study integrated textual data with satellite imagery. While multimodal inputs improved accuracy, challenges remained in extracting precise and reliable insights.

While these studies demonstrate the potential of AI for processing text and images, video content—a medium that can provide rich and detailed accounts of disaster impacts (Schlein et al., 2024)—remains largely underutilized. This gap presents an opportunity to extend the capabilities of AI beyond text and images to include MLLMs. MLLMs can process and integrate multiple data types—text, images, and videos—for a more comprehensive analysis of social media content. While numerous applications for multimodal data have been proposed, researchers have noted ongoing concerns around computing power (Li et al., 2024a).

Although some recent studies have begun to explore multimodal approaches (Andreadis et al., 2023; Koshy & Elango, 2023), there is little work specifically addressing the use of MLLMs to analyze video content for emergency management. This paper builds on this emerging body of work by testing the feasibility of using a state-of-the-art MLLM—Gemini 1.5 Flash—for filtering social media videos shared during hurricanes. Unlike prior work that focuses

on controlled datasets or theoretical frameworks, our study evaluates prompting strategies using real-world data from X (formerly Twitter). In doing so, we contribute empirical insights on the use of MLLMs for video classification in time-sensitive, high-stakes contexts such as emergency response.

METHODOLOGY

This study explores how MLLMs can be used to identify user-generated videos that contain relevant information shared on social media during disaster events. In our initial consultations with emergency managers, they expressed interest in using MLLMs to find videos that showed damage to key infrastructure, such as electricity, water, and roads. However, we determined that a narrower focus was necessary for testing given the complexity of the task. Hurricanes were selected as the primary use case due to their distinct visual features—such as high winds, heavy rain, storm surges, and infrastructure damage—that are often captured in disaster footage.

Our initial tests, informed by prior research, revealed the MLLMs struggled with questions directly related to emergency management processes, such as identifying infrastructure damage. The models frequently produced inaccurate or unreliable responses. As a result, we shifted our approach to first filtering videos to include only those showing dangerous weather conditions, thereby reducing the volume of irrelevant content that emergency managers would need to review manually. Like the findings of (Otal et al., 2024), we found that there were many irrelevant videos, such as memes or unrelated content. Additionally, many videos that contained hurricane-related primarily featured weather maps, which emergency managers already had access to through official sources. Filtering out irrelevant videos became an important step toward streamlining the identification of useful disaster-related footage.

A primary goal of this study was to evaluate the influence of various prompting strategies and question styles on the quality and utility of MLLM responses for these emergency managers. To achieve this, we collected videos from social media, focusing on disaster-related hashtags for hurricanes. These videos were then processed through an MLLM, which provided a chat interface for posing questions about each video and receiving responses. We systematically evaluated different prompting strategies, recorded the MLLM's outputs, and assessed their accuracy. This approach allowed us to find the most effective approaches for identifying relevant video data in disaster scenarios. Below, we detail the key components of the research, including data collection, model selection, prompting strategies, and evaluation techniques.

Data Collection

Hurricanes were selected as the disaster type for this study due to their well-documented and visually distinct characteristics. Three hurricanes—Ian, Laura, and Dorian—were chosen based on the availability of publicly shared datasets containing video content. From each hurricane dataset, videos were filtered to include only those ending in the .mp4 format. To manage input constraints for MLLMs, longer videos were trimmed to one minute. A total of 50 videos per hurricane were randomly sampled, with efforts made to minimize redundancy across the selected clips.

Model Selection

The initial proof of concept was conducted using the VideoChat2 MLLM (Li et al., 2024b), because it had a publicly available interface that could process video inputs. However, limitations in speed and the subsequent unavailability of the model led to the adoption of Gemini Flash 1.5. This publicly available and frequently updated model provided the necessary efficiency, performance, and API support required to conduct this study.

Prompting Strategies

To filter out irrelevant videos, such as memes and weather maps, we designed prompts focusing on three key criteria: whether the video contained a weather map, visible evidence of a storm, and indications of danger. Videos that met the latter two criteria while excluding weather maps were classified as relevant for emergency managers.

The benchmark that (Rawat, 2024) created for LLM accuracy in disaster management provided some valuable insights into effective prompting strategies. Most of the models that Rawat used struggled with open-ended questions, so our approach prioritized testing the accuracy of the more reliable closed-question formats. The specific prompting styles were selected based on preliminary tests and focused on methods that addressed known limitations of MLLMs. For

example, few-shot and multi-shot prompting techniques were excluded since they require prior knowledge of the events depicted in the video, which is often impractical for real-time disaster response scenarios

Prompts were divided into two primary groups: those limited to closed questions and those accommodating both open-ended and closed questions. Closed questions require binary or constrained responses (e.g., yes/no or multiple choice), while open-ended questions demand reasoning and longer responses. The tables below show the prompting styles used in this study and specify whether they support only closed questions or a combination of both question types.

Table 1. Closed Only Prompting Strategies

Strategy	Description	Example
Multiple Choice	Provides structure to the response by giving the prompt a limit to what answers it can give. The VideoChat paper included this method of prompting to show how developed their model is (Li et al., 2024b).	This video shows: a) Weather Map b) Ninjas c) Dinosaurs d) None of the Above
Generated Knowledge	Adds a fact to the prompt to give the model additional knowledge that it may not have already possessed (Liu et al., 2022).	A weather map is any map or chart that shows the meteorological elements at a given time over an extended area. Is this video of a weather map?

Multiple Choice and Generated Knowledge (see Table 1) were not conducive to open-ended questions, as introducing additional reasoning would alter their structured nature. For example, it is difficult to develop multiple choice answer options for open-ended questions. Generated knowledge provided a fact about general information, but there were not facts that could be helpful about open-ended questions which referred to specifics about the video. The prompting strategies in Table 2 were more conducive to reasoning-based responses, which made them suitable for evaluating both closed and open-ended question styles.

The last variable introduced into the prompting styles examines the impact of context retention using two key functions from Gemini's API¹. The *generate_content* function serves as the standard query method, where the model's context window is cleared between questions. This prevents it from retaining knowledge from prior interactions. In contrast, the *start_chat* function allows for multiple rounds of questions and responses that build upon one another. This function facilitates a step-by-step progression toward answers that lets users address multipart problems or refine queries iteratively. With *start_chat* prompting, the sequence of questions matters because previous questions and responses influence subsequent responses. This approach allows for more contextual and logically connected dialogue, which we hypothesize could improve the response accuracy in disaster-related video analysis.

Table 2. Both Closed and Open-Ended Prompting Strategies

Strategy	Description	Example
Control	Simulates the way that basic users may ask questions.	Is this video of a weather map?
Chain of Thought	Involves adding the statement "Let's think step by step" to the end of a question in an attempt to encourage reasoning through explanation (Wei et al., 2022).	Is there danger in this video that would hinder the response of emergency managers? Let's think step by step
Role Based	Invites the model to answer the question how it would if it had a certain job or role (Kong et al., 2024).	Imagine you are a weatherman looking at this video. Does it show a storm?

To assess the impact of context retention on response quality, both cleared and uncleared styles were tested across all prompting strategies. The cleared style refers to interactions using the *generate_content* function, where the model does not retain prior exchanges. The uncleared style refers to interactions using the *start_chat* function, where the model maintains conversational history. Throughout the rest of this paper, we use the terms *cleared* and *uncleared* to

¹ <https://ai.google.dev/gemini-api/docs>

distinguish between these styles. This comparison allowed for an evaluation of whether retaining context improves the accuracy and relevancy of MLLM responses in the context of disaster video analysis.

Table 3. Evaluation Criteria for Prompt Responses and Examples

	Question	Answer	Video Description
Correct (1 point)	This video shows: a) Weather Map b) Ninjas c) Dinosaurs d) None of the Above	The answer is a). The video shows a weather map of a hurricane.	The video shows a repeat scene of a weather map, with the hurricane approaching Texas and Louisiana
Partially correct or inconclusive (0.5 points)	Does the video show weather conditions that would be dangerous for people in that area?	It is not possible to tell from the short video if there are any dangerous weather conditions in the area. The video is too short, and doesn't show enough detail to determine weather conditions.	The video shows a group of people setting up a mobile weather station in the middle of the night.
Incorrect or irrelevant response (0 points)	Does the video show weather conditions that would be dangerous for people in that area? Let's think step by step	Let's think step by step: 1. What is the video about? The video shows a group of people dancing and singing. 2. What weather conditions are shown in the video? There are flames in the video. 3. Are flames dangerous to people? Yes, flames can cause burns and other injuries. Therefore, the video shows a weather condition (flames) that would be dangerous for people in that area.	This video is from a K-pop music video. There are people dancing with a fire behind them.

Evaluation

After receiving API responses for each video and prompting strategy, accuracy was assessed using a three-tier grading scale (see Table 3). Each response was manually reviewed and scored based on alignment with the evaluation criteria. The total score for each prompt style was calculated by summing individual evaluations across all videos. To determine an overall accuracy measure for each prompt style, the total score was divided by the number of videos to provide an average accuracy score.

RESULTS

In this section, we present the results of evaluating different prompt strategies for filtering hurricane-related videos generated by the public using the Gemini Flash 1.5 model. The evaluation is organized around the three types of questions we asked about weather maps, storms, and danger. For each category, a total of 150 videos were tested. A perfect MLLM would produce a score of 150. The final accuracy score for each question type was calculated by dividing the total score by 150 to determine the average performance. Each subsection reports on and discusses the strengths and weaknesses of the prompting techniques in achieving accurate, relevant, and actionable responses for emergency management.

Weather Map

Weather maps were anticipated to be the simplest category for the MLLM to analyze. Videos of weather maps are frequently shared during hurricanes but are often redundant and uninformative for emergency managers, who already have access to this information from official sources like the National Weather Service in the US. Thus, correctly identifying and filtering these videos could be helpful in reducing noise in datasets, as they tend to represent a disproportionate amount of the content found on social media.

Table 4 summarizes the accuracy scores for different prompting styles and question types in detecting weather maps. The results reveal consistently high performance across all styles, with most achieving an accuracy of 0.9 (out of 1) or higher. The highest accuracy score (1.0) was for the *Role-Based Uncleared Closed* prompt, where the model was instructed to take on a specific role: *“Imagine you are a weatherman looking at this video. Does it show a weather*

map?” This suggests that aligning the model’s role with the task can improve accuracy, particularly for clear and structured questions.

Table 4. Weather Map Prompt Accuracy Scores

Prompting Style	Question Style	Uncleared		Cleared	
		Score (0-150)	Average (0-1)	Score (0-150)	Average (0-1)
Multiple Choice	Closed	149.0	0.99	145.5	0.97
Generated Knowledge	Closed	149.0	0.99	149.0	0.99
Control	Closed	143.0	0.95	140.0	0.93
	Open-ended	137.5	0.92	136.5	0.91
Chain of Thought	Closed	149.0	0.99	148.5	0.99
	Open-ended	137.0	0.91	140.0	0.93
Role Based	Closed	150.0	1.00	142.5	0.95
	Open-ended	135.0	0.90	132.0	0.88

Open-ended questions performed slightly worse in all prompting styles, which aligns with expectations. These questions require broader reasoning and leave more room for interpretation and possible error (e.g., Does the video show weather conditions that would be dangerous for people in that area?). The model particularly struggled when asked open-ended questions about dangerous weather conditions shown in the map rather than simply identifying a weather map itself. The highest score for open-ended questions was achieved using *Chain of Thought* prompting in both for uncleared and cleared conditions. Given that *Chain of Thought* is designed to simulate advanced reasoning, this suggests it may be one of the most effective prompting styles for handling more complex queries.

Another finding is the performance disparity between the cleared and uncleared conditions. Cleared prompts generally resulted in lower scores, as removing prior context seemed to reduce the model’s ability to use accumulated information. However, exceptions like *Generated Knowledge* and *Chain of Thought Open-ended* prompts showed that some strategies benefit from a context reset, particularly when structured reasoning is emphasized.

Overall, these results suggest that closed questions, role-based prompting, and Chain of Thought techniques yield the most reliable results for identifying weather map videos. However, additional refinement may be needed for open-ended queries to improve accuracy in more complex disaster-related assessments.

Storm

Storm-related videos, which depict phenomena like rain, wind, and storm surge, proved more challenging for the MLLM to classify accurately. Initial experiments attempted to break down storm features into separate questions (e.g., distinguishing between rain and wind), but this approach led to more inaccuracy. Consequently, we refined our analysis to focus on holistic questions about storm presence rather than individual weather aspects.

Table 5 presents the accuracy scores for storm-related prompts. Overall, performance was lower than for weather maps, with scores ranging from 0.88 to 0.95. This result suggests that storm classification introduces additional complexity and subjectivity, as visual cues such as cloud movement or water levels are not always definitive indicators of storm presence.

Among prompting styles, *Generated Knowledge* achieved the highest accuracy scores for both cleared and uncleared conditions. Providing the model with a fact—e.g., “Thunderstorms include thick clouds, heavy rain or hail, lightning, thunder, and strong winds...”—seemed to enhance its reasoning about storm characteristics. Notably, responses often included statements acknowledging that while the video did not depict a textbook thunderstorm, it still contained

storm conditions. This would seem to indicate that knowledge augmentations can help improve a model’s interpretive capabilities.

Table 5. Storm Prompt Accuracy Scores

Prompting Style	Question Style	Uncleared		Cleared	
		Score (0-150)	Average (0-1)	Score (0-150)	Average (0-1)
Multiple Choice	Closed	138.0	0.92	138.5	0.92
Generated Knowledge	Closed	142.0	0.95	140.0	0.93
Control	Closed	139.0	0.93	143.5	0.96
	Open-ended	132.0	0.88	135.0	0.90
Chain of Thought	Closed	139.0	0.93	135.5	0.90
	Open-ended	140.0	0.93	134.0	0.89
Role Based	Closed	141.5	0.94	142.0	0.95
	Open-ended	140.0	0.93	132.0	0.88

The results for cleared prompts in storm-related questions were mixed, revealing differences in how context retention impacts accuracy. For instance, the *Control Closed Cleared* prompt scored higher than its uncleared counterpart (0.96 vs 0.93). This suggests that clearing context can sometimes mitigate the compounding effect of initial errors. This is particularly useful for straightforward, factual evaluations where prior responses might introduce bias. However, for reasoning-intensive styles like *Chain of Thought Open-ended*, uncleared prompts outperformed cleared ones. Retaining context in these cases allowed the model to build on earlier responses, demonstrating the benefits of accumulated reasoning for more complex tasks.

Interestingly, cleared prompts performed better overall for storms than they did for weather maps, with exactly half of the storm-related prompting styles achieving higher scores when context was reset. One notable drawback of uncleared prompts is their tendency to persist with an initial answer throughout a sequence of related questions, even when that answer is incorrect. For example, if the model incorrectly identified a storm in the initial question, it often maintained this misclassification across subsequent prompts. This points to the possible trade-off between leveraging context for iterative reasoning and the risk of perpetuating errors when context is retained.

These findings show that while structured knowledge prompts can improve storm classification accuracy, the choice between cleared and uncleared context should be carefully considered based on the specific task. For more objective, fact-based assessments, clearing context appears beneficial, while for reasoning-heavy tasks, maintaining context may offer advantages—provided errors are managed effectively.

Danger

Danger-related questions were expected to yield the lowest accuracy due to their inherent subjectivity and the challenge of identifying actionable threats from video content. The concept of danger is highly context-dependent and often requires nuanced reasoning that balances visual cues with broader situational knowledge. This complexity led to misclassifications such as the example seen in Table 3, where the MMLM incorrectly categorized fire in a music video as a dangerous weather condition.

Table 6 shows the accuracy scores for danger-related prompts, which had the most variance, ranging from 0.57 to 0.92. This variability reflects the subjectivity involved in assessing danger in videos. The lowest score (0.57) was observed for *Multiple Choice Closed Cleared* prompts, where limiting the response options seemed to hinder the model’s ability to evaluate complex situational threats. Context retention played a crucial role in the accuracy for these cleared responses. In uncleared conditions, the model could infer from prior exchanges that evaluation focused on

weather-related dangers. However, when the context was cleared between prompts, the model often defaulted to assuming that most situations were dangerous, even when no clear weather-related threat was present.

Table 6. Danger Prompt Accuracy Scores

Prompting Style	Question Style	Uncleared		Cleared	
		Score (0-150)	Average (0-1)	Score (0-150)	Average (0-1)
Multiple Choice	Closed	131.0	0.87	85.0	0.57
Generated Knowledge	Closed	134.0	0.89	130.0	0.87
Control	Closed	133.0	0.89	130.0	0.87
	Open-ended	138.5	0.92	109.0	0.73
Chain of Thought	Closed	132.0	0.88	138.5	0.92
	Open-ended	133.5	0.89	104.5	0.70
Role Based	Closed	133.5	0.89	134.0	0.89
	Open-ended	135.5	0.90	121.5	0.81

Similarly, *Chain of Thought Open-ended Cleared* performed poorly because the model frequently overgeneralized danger without prior context to ground its reasoning in weather-related circumstances. Conversely, the highest score (0.92) was observed for *Chain of Thought Closed Uncleared* prompts, suggesting that step-by-step reasoning, without a context reset, is particularly effective for this complex classification task. These results indicate that closed questions seem to be more effective when filtering danger-related content, as they provide clearer parameters for classification.

One challenge in evaluating danger-related responses was the subjectivity involved in grading accuracy. To minimize bias, danger was only marked as "correct" if it was directly related to weather conditions. For example, a video of a fire was deemed irrelevant unless it was caused by weather phenomena. However, the MLLM sometimes misinterpreted danger based on unrelated factors. For instance, in one response, the model stated: "This video depicts a young child using foul language, which is not safe... However, there is no evidence of immediate danger that would hinder emergency management. It's not appropriate for a child to use such language, and it's important to address the issue with the child's parents or guardians." This response illustrates the difficulty of aligning the model's reasoning with specific emergency management priorities, as it misclassifies social risks as equivalent to physical threats.

Generally cleared prompts performed worse than uncleared prompts for danger detection, with only one out of eight prompting styles resulting in a higher score than its uncleared alternative. This suggests that retaining contextual information is important for reasoning-intensive tasks like danger evaluation, where the model must build upon prior responses to form accurate conclusions.

Overall Performance

Table 7 summarizes the weighted scores across all three question categories (weather map, storm, and danger). These results show several trends.

Uncleared prompts generally outperformed cleared prompts, particularly for reasoning-intensive tasks like storm and danger detection. Retaining context allowed the model to build on prior responses, improving its ability to reason iteratively and refine answers across related questions. However, this advantage came with a risk of perpetuating initial errors when incorrect conclusions were drawn early in the sequence.

Prompting styles that simulate reasoning, such as *Chain of Thought* and *Generated Knowledge*, demonstrated the highest overall accuracy. These strategies equipped the model with tools to interpret complex scenarios by providing step-by-step reasoning or supplemental facts. Their effectiveness seems to show the importance of designing prompts that encourage logical progression and structured reasoning.

Table 7. Total Prompt Accuracy Scores

Prompting Style	Question Style	Uncleared		Cleared	
		Score (0-150)	Average (0-1)	Score (0-150)	Average (0-1)
Multiple Choice	Closed	419.5	0.93	372.0	0.83
Generated Knowledge	Closed	424.5	0.94	421.0	0.94
Control	Closed	419.0	0.93	412.0	0.92
	Open-ended	415.5	0.92	378.0	0.84
Chain of Thought	Closed	423.5	0.94	427.0	0.95
	Open-ended	417.5	0.93	377.5	0.84
Role Based	Closed	429.0	0.95	413.5	0.92
	Open-ended	421.5	0.94	387.0	0.86

Closed questions outperformed open-ended ones across nearly all categories, particularly for straightforward classification tasks like weather maps and storms. Open-ended prompts introduced ambiguity, making it harder for the model to extract concise and accurate answers. For tasks requiring factual precision, closed questions proved to be a more reliable approach.

Role-Based prompts excelled in clear-cut, structured tasks but struggled with more subjective evaluations like danger detection. Aligning the model's "role" with the task at hand—such as asking it to imagine being a weatherman—helped improve performance for specific scenarios. However, this approach required refinement when applied to more complex reasoning tasks, where additional context and flexibility were needed.

The highest overall score (0.95) was achieved by the *Closed Chain of Thought Cleared* prompts. This configuration balanced logical reasoning with the benefits of a context reset, ensuring high accuracy without being influenced by prior responses. This points to the importance of crafting prompt strategies that combine structured reasoning with mechanisms to mitigate bias or compounding errors.

DISCUSSION

This research represents an initial step toward leveraging MLLMs to classify disaster-related videos on social media, with the goal of improving situational awareness for emergency responders. By evaluating the Gemini Flash 1.5 model, we found that prompting style significantly affects classification performance. These findings highlight the need for tailored prompt engineering when using MLLMs in real-time disaster contexts.

Our results demonstrate that techniques such as context retention, Chain of Thought prompting, and knowledge augmentation can improve the accuracy of video classification. Retaining conversational history enabled iterative refinement for complex tasks like storm and danger detection. However, it also introduced a risk of error propagation when the model misclassified content early in the sequence. Structured reasoning approaches like Chain of Thought were particularly effective for multi-step evaluation tasks, while knowledge augmentation helped the model contextualize ambiguous visual cues. The performance differences across these prompting strategies indicate that a one-size-fits-all approach is insufficient for filtering disaster-related multimedia content. Instead, adaptive prompting strategies tailored to specific classification tasks are necessary to optimize performance.

Given the rapid evolution of MLLM technology, it is impractical to rely on a single filtering method that may become obsolete in future model updates. Instead, this research emphasizes the need to develop adaptable evaluation frameworks to evaluate and refine MLLM performance across diverse disaster scenarios. Practitioners can use the framework provided here to test and optimize prompts for their specific use cases.

Finally, although this study focused on hurricanes, the methodology could be extended to other types of disasters, such as wildfires, floods, or blizzards. Each introduces unique classification challenges—for example, distinguishing between wildfire smoke and regular fog or assessing the severity of flooding based on water depth. Our approach can support task-specific prompt engineering across these diverse contexts.

Limitations & Future Directions

This study presents initial findings on the use of MLLMs for filtering disaster-related videos, but several limitations point to opportunities for future research. First, the dataset was relatively small and restricted to one-minute video clips due to Gemini’s input constraints. This may have excluded important context about evolving disaster conditions. Future work should explore processing longer videos and consider converting non-MP4 files rather than discarding them to expand dataset size and diversity.

Second, our evaluation relied on manual human review, which, while guided by standardized rubrics, may introduce subjectivity—particularly in tasks such as danger detection. Incorporating ground-truth labels from domain experts or using automated evaluation metrics could improve consistency and reliability. Additionally, prompt sequencing influenced results in some conditions where prior responses were retained, occasionally leading to error propagation. Future studies could test different prompt sequencing strategies or introduce mechanisms to selectively reset context.

Finally, this study focused exclusively on Gemini Flash 1.5 and did not compare MLLM performance with other approaches, such as deep learning-based video classifiers or other LLM architectures like GPT or LLaMA (though currently they lack video support). Expanding the evaluation across models and baseline methods would strengthen the generalizability of our findings. Similarly, while we focused on hurricane-related content from X, future work could apply this framework to other disaster types and platforms like Instagram or Facebook, which may reflect different demographics or types of content. As MLLMs evolve, continued research will be needed to adapt evaluation strategies and ensure that these models remain viable tools for real-time crisis response.

CONCLUSION

This study provides an initial exploration into the use of MLLMs for disaster classification and offers insights into how different prompting strategies impact performance. While context retention, structured reasoning, and knowledge augmentation emerged as factors in improving classification accuracy, their effectiveness varies based on the nature of the task. The findings point to the importance of prompt engineering and adaptive evaluation methods as MLLM capabilities continue to evolve. Future efforts should focus on scaling these methods to broader disaster scenarios, incorporating real-time processing, and refining evaluation techniques to further enhance the reliability of AI-driven disaster response tools.

REFERENCES

- Andreadis, S., Bozas, A., Gialampoukidis, I., Moutmtzidou, A., Fiorin, R., Lombardo, F., Mavropoulos, T., Norbiato, D., Vrochidis, S., Ferri, M., & Kompatsiaris, I. (2023). DisasterMM: Multimedia Analysis of Disaster-Related Social Media Data Task at MediaEval 2022. *MediaEval*.
- Barman, D., & Colan, O. (2023). Does Explanation Matter? An Exploratory Study on the Effects of Covid-19 Misinformation Warning Flags on Social Media. *2023 10th International Conference on Behavioural and Social Computing (BESC)*, 1–7. <https://doi.org/10.1109/BESC59560.2023.10386371>
- Colverd, G., Darm, P., Silverberg, L., & Kasmanoff, N. (2023). FloodBrain: Flood Disaster Reporting by Web-based Retrieval Augmented Generation with an LLM. <https://doi.org/10.48550/ARXIV.2311.02597>
- Fang, J. (2024). Optimizing AI-Driven Disaster Management through LLMs. <https://doi.org/10.20944/preprints202407.1446.v1>
- Freeman, C., Nunnari, N., Edgemon, L., & Marsh, K. (2021). Improving Public Messaging for Evacuation and Shelter-in-Place. *FEMA*. https://www.fema.gov/sites/default/files/documents/fema_improving-public-messaging-for-evacuation-and-shelter-in-place_literature-review-report.pdf
- Castro Herrera, L. (2021). Configuring Social Media Listening Practices in Crisis Management. *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 2021)*. http://idl.iscram.org/files/luciacastroherrera/2021/2361_LuciaCastroHerrera2021.pdf

- Hiltz, S. R., Hughes, A. L., Imran, M., Plotnick, L., Power, R., & Turoff, M. (2020). Exploring the usefulness and feasibility of software requirements for social media use in emergency management. *International Journal of Disaster Risk Reduction*, 42, 101367. <https://doi.org/10.1016/j.ijdr.2019.101367>
- Imran, M., Alam, F., Qazi, U., & Peterson, S. (2020). Rapid Damage Assessment Using Social Media Images by Combining Human and Machine Intelligence. *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 2020)*. https://idl.iscram.org/files/muhammadimran/2020/2269_MuhammadImran_etal2020.pdf
- Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys*, 47(4), 67:1-67:38. <https://doi.org/10.1145/2771588>
- Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024). Better Zero-Shot Reasoning with Role-Play Prompting. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4099–4113. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.228>
- Koshy, R., & Elango, S. (2023). Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, 35(2), 1607–1627. <https://doi.org/10.1007/s00521-022-07790-5>
- Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., & Qiao, Y. (2024a). VideoChat: Chat-Centric Video Understanding. <https://doi.org/10.48550/arXiv.2305.06355>
- Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Lou, P., Wang, L., & Qiao, Y. (2024). MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22195–22206. <https://doi.org/10.1109/CVPR52733.2024.02095>
- Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Le Bras, R., Choi, Y., & Hajishirzi, H. (2022). Generated Knowledge Prompting for Commonsense Reasoning. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3154–3169. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.225>
- Otal, H. T., Stern, E., & Canbaz, M. A. (2024). LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration. *Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI)*, 851–859. <https://doi.org/10.1109/CAI59869.2024.00159>
- Palen, L., & Hughes, A. L. (2018). Social Media in Disaster Communication. In H. Rodríguez, W. Donner, & J. E. Trainor (Eds.), *Handbook of Disaster Research* (pp. 497–518). Springer International Publishing. https://doi.org/10.1007/978-3-319-63254-4_24
- Purohit, H., Buntain, C., Hughes, A. L., Peterson, S., Lorini, V., & Castillo, C. (2025). Engage and Mobilize! Understanding Evolving Patterns of Social Media Usage in Emergency Management. *Proceedings of the ACM on Human-Computer Interaction, CSCW*. <https://doi.org/10.1145/3710965>
- Rawat, R. (2024). DisasterQA: A Benchmark for Assessing the performance of LLMs in Disaster Response. <https://doi.org/10.48550/ARXIV.2410.20707>
- Reuter, C., Hughes, A. L., & Kaufhold, M.-A. (2018). Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human-Computer Interaction*, 34(4), 280–294. <https://doi.org/10.1080/10447318.2018.1427832>
- Samadi, V., Stephens, K. K., Hughes, A., & Murray-Tuite, P. (2024). Challenges and opportunities when bringing machines onto the team: Human-AI teaming and flood evacuation decisions. *Environmental Modelling & Software*, 175, 105976. <https://doi.org/10.1016/j.envsoft.2024.105976>
- Schlein, A., Wang, S., Remaker, V., Tie, Z., Haughey, M. M., Davidson, R. A., Kendra, J., & Starbird, K. (2024). Bursting Pipes and Boiling Snow: Disaster Impacts and Adaptations in the 2021 Texas Power Crisis from the Lens of Short-Form Social Media Videos. *Journal of Disaster Studies*, 1(2), 220–248.
- Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. *iConference 2014 Proceedings*. <https://doi.org/10.9776/14308>
- Sutton, J., Palen, L., & Shklovski, I. (2008). Backchannels on the Front Lines: Emergent Uses of Social Media in the

- 2007 Southern California Wildfires. *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 2008)*. http://idl.iscram.org/files/sutton/2008/985_Sutton_etal2008.pdf
- Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1079–1088. <https://doi.org/10.1145/1753326.1753486>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837.
- Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., & Starbird, K. (2018). From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 195:1-195:18. <https://doi.org/10.1145/327446>