

Name: Md Mehedi Hasan Mixon

ID: 2436061

Supervisor: Anum Paracha

**EVALUATING THE RELIABILITY OF GENERATIVE AI
FOR DETECTING FALSE AND MISLEADING PUBLIC-
SAFETY INFORMATION DURING EMERGENCY EVENTS**

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who helped me through the completion of this dissertation. I am very grateful to my supervisor for his helpful tips, constructive criticism, and ceaseless encouragement throughout the research process. I would also like to thank my lecturers and staff for supporting me throughout my coursework. I thank my family and friends for their patience, motivation and understanding throughout the lengthy process of completing this work. Last but not least, I would like to express my gratitude to all those who helped with this research or contributed to it indirectly. Thanks to their encouragement and support, I was able to successfully complete this dissertation.

ABSTRACT

This dissertation assesses the reliability of Generative Artificial Intelligence (AI) in identifying inaccurate and misleading public-safety information during emergencies. It looks beyond simple accuracy, examining whether the model can distinguish three categories: Reliable/Supported, Misleading/False and Unverified. The research is based on the PHEMEPlus dataset, from which 1,972 rumours were extracted across five emergency events – Charlie Hebdo, Ferguson, Germanwings, the shooting in Ottawa, and the siege in Sydney. For the experimental evaluation, a balanced sample of 300 records comprising 100 from each category was selected. The study uses the openai/gpt-oss-120b model with three prompting strategies: Zero-Shot, Definition-Based and Few-Shot prompting, and each condition is designed to be repeated three times to assess consistency and prompt sensitivity. The evaluation framework is quantitative, covering accuracy, precision, recall, Macro-F1, false-positive rate, false-negative rate and confusion matrices, and qualitative, covering hallucinations, unsupported explanations, forced classifications and safety-related misidentifications. A key strength of the research is that it keeps a separate Unverified category, so that claims lacking evidence are not automatically treated as false. Because complete performance metrics were not available at the time of writing, the results concern the dataset, the evaluation framework and pilot behaviour rather than final accuracy figures. Based on these findings and the evaluation framework, Generative AI can be an effective complement to information processing and the initial triage phase, but should not be used as a definitive decision-making tool in the event of an emergency with high risks. When information is not complete or if there is a risk of classification errors that may have a serious impact, human verification is essential. The research also emphasises the need for repeatable experimental configurations, careful prompt design, error analysis and auditability in assessing Generative AI for public safety applications.

Table of Contents

Chapter 1: Introduction	6
1.1 Background of the Study	6
1.2 Generative AI and Emergency Information Verification	6
1.3 Problem Statement	7
1.4 Research Aim	8
1.5 Research Objectives	8
1.6 Research Questions	9
1.7 Significance of the Study	9
1.8 Chapter Summary	10
Chapter 2: Literature Review	11
2.1 Overview of Literature Review	11
2.2 Public-Safety Information During Emergency Events	11
2.3 False, Misleading and Disinformation in Emergencies	12
2.4 Emergency Events and the Spread of Misinformation	14
2.5 Social Media and Digital Platforms as Information Channels	16
2.6 Generative AI and Large Language Models for Classification	16
2.7 Reliability, Accuracy and Consistency of Generative AI	18
2.8 Hallucination, Unsupported Explanations and AI-Generated Errors	19
2.9 Human Verification and Human-in-the-Loop Decision Support	21
2.10 Critical Review of Existing Research	22
2.11 Research Gap	23
2.12 Conceptual Framework	23
2.13 Chapter Summary	24
Chapter 3: Methodology	26
3.1 Research Approach	26
3.2 Research Design	26

3.3 Research Strategy	27
3.4 Research Method.....	28
3.5 Data Collection and Analysis	29
3.6 Validity and Reliability	30
3.7 Ethical Consideration	31
3.8 Summary	32
Chapter 4: Results and Analysis	33
4.1 Introduction	33
4.2 Implementation and Evaluation	33
4.3 Key Findings	44
4.4 Critical Analysis.....	46
4.5 Summary	48
Chapter 5: Conclusion and Recommendations	49
5.1 Conclusion.....	49
5.2 Recommendations	50
5.3 Limitations of the Study	50
5.4 Practical Implications.....	51
5.5 Future Research.....	51
5.6 Final Summary	52
References.....	53

Chapter 1: Introduction

1.1 Background of the Study

The rapid growth of social media has dramatically changed the way information is generated, shared and accessed during disasters (Miller, 2024). During floods, wildfires and earthquakes, social media platforms carry first-hand reports, local public-safety information, requests for assistance, and updates from government, affected communities and the media. These data can play a significant role in situational awareness and humanitarian information management since emergency information can be exchanged quickly and widely among a large population (Komendantova and Erokhin, 2025). But the same qualities that make social media a useful tool in times of emergency can pose significant dangers. The sheer speed, availability and volume of online communications may contribute to the fast spread of rumours, unverifiable reports, misinformation and false information.

In a crisis, information sources become even more critical because decisions may have to be made rapidly without perfect information. Information from social media can help members of the public decide whether to evacuate, seek help, avoid certain areas or act on warnings (Truong, 2026). Online information can also help emergency responders and analysts to make sense of the unfolding situation. Therefore, misinformation can misrepresent an emergency, affect public actions and lead to lost trust in authorities. However, information that seems uncertain should not be rated false simply because supporting evidence is not yet available, since some early emergency reports later prove accurate (Guo et al., 2023). This leaves a difficult balance to be struck between trustworthy information, misleading or false information, and unverified information.

Social media has proven helpful in comprehending emergencies and identifying humanitarian information, as shown by research in crisis informatics (Altayeva, 2025). Existing datasets such as CrisisMMD, HumAID and CrisisBench can be used to analyse emergency-related information. But information veracity and emergency relevance are two different things. A post can be very well targeted to an emergency and contain incorrect, incomplete, out-of-date or unverifiable information. It therefore cannot be assumed that systems which identify useful emergency information can also discern whether a claim is true or false.

1.2 Generative AI and Emergency Information Verification

Generative Artificial Intelligence (AI), particularly Large Language Models (LLMs), provides a potentially useful approach for analysing large quantities of textual information. LLMs can provide zero-shot as well as few-shot classification using language prompts, which are a new approach to classification without the need for a specifically trained classifier (Chowdhury, 2026). This capability makes it possible to analyse social media content quickly at the start of an emergency response and can help analysts identify potentially problematic information.

However, a Generative AI system's capacity to create coherent and believable answers is not necessarily an indicator that its choices are accurate. LLM outputs can be affected by the wording, structure and complexity of prompts. Semantically similar prompts, different definitions, or repeated evaluation of the same information can all produce different answers. Models can also produce hallucinations or explanations that lack sufficient evidence (Ibrahim et al., 2026). A model's stated confidence is also not a direct measure of its accuracy. When this is put into the context of public-safety applications, where a misjudgment could have life-and-death implications, these limitations are especially important.

Therefore, the measurement of Generative AI reliability in the analysis of emergency information needs to involve a wider range of measurements than classification accuracy (Jambholkar, 2024). Although a system might be accurate overall, it may perform poorly in specific safety-related categories. There are also false negatives and false positives to consider. Treating inaccurate information as genuine may help harmful information spread, whereas mistaking accurate emergency advice for false information may cause it to be ignored. Likewise, categorising an unverified claim as true or false without further evidence could create a false sense of certainty.

1.3 Problem Statement

The main question of this dissertation is whether generative AI is capable of accurately identifying misinformation and verified public-safety information in social media postings during emergencies. Some existing crisis social-media tools categorise information based on whether it is informative or humanitarian. The classifications are useful when conducting information triage, but not necessarily when determining whether a claim is true or false (Shoaib et al., 2023). Using humanitarian classifications as measures of veracity therefore raises problems of construct validity.

Rumour-verification work also suggests that it is sometimes necessary to gather evidence and conversational context more widely than just from the content of a single message to establish online rumours' truth. Concurrently, recent studies on LLM classifications reveal the potential effect of prompt phrasing and prompt complexity on classification results (Wang et al., 2026). The research problem is therefore not only how accurately Generative AI classifies emergency posts. It is also whether the system makes consistent and sound judgements when the same data are processed several times, and when the prompt wording varies without changing what is being asked.

The dissertation therefore proposes a controlled experiment that uses a Python pipeline to classify emergency-related rumours. Three categories are given operational definitions: Reliable/Supported, Misleading/False and Unverified. A variety of prompting strategies such as zero-shot, definition-based and few-shot prompting will be tested. The evaluation examines accuracy, Macro-F1, precision, recall, false-positive rate, false-negative rate, confusion matrices, consistency and prompt sensitivity. Particular attention is also paid to hallucinations, unsupported explanations, forced classifications, and safety-related misidentifications.

1.4 Research Aim

The aim of this dissertation is:

To assess the reliability of Generative AI in detecting and categorising misinformation on safety in emergency-related social media postings, with a focus on the model's performance, consistency, prompt sensitivity, handling of uncertainty and errors.

1.5 Research Objectives

The research will accomplish this aim through the following objectives:

1. To develop a reproducible Generative AI evaluation framework for classifying emergency-related information as Reliable/Supported, Misleading/False, or Unverified.
2. To evaluate and compare zero-shot, definition-based, and few-shot prompting using classification performance, consistency, and prompt sensitivity.
3. To analyse Generative AI errors, including hallucinations, unsupported explanations, forced classifications, and safety-related misidentifications.

1.6 Research Questions

The following main research question drives the dissertation:

How well can Generative AI distinguish misinformation from verified public-safety information?

The questions that support the research question are:

- What level of accuracy and class-level performance can be achieved by the chosen Generative AI model?
- How often does the model produce false-positive and false-negative classifications?
- How impactful is the prompt design on classification performance?
- Does the model give consistent responses when the same information is analysed repeatedly?
- Does Generative AI give a different response when the prompt wording changes but its meaning does not?
- Which claims about public safety are more challenging for Generative AI to classify?
- In which situations should Generative AI defer to human verification?

1.7 Significance of the Study

The study is valuable because it goes beyond previous research that tested only whether Generative AI can detect emergency-related content, and considers whether the system can support responsible decisions that are accurate, repeatable and appropriately cautious. The research is particularly relevant to the public-safety threats that arise from emergency misinformation, and to those that arise when accurate emergency information is rejected by the public.

This study provides an evaluation framework that combines crisis informatics, misinformation verification and Generative AI reliability. Its results will help demonstrate the strengths and weaknesses of various prompting strategies, and determine when human verification is

required. Practical safeguards covered by this framework include explicit labelling rules, systems with humans in the loop on high-risk cases, abstention options and audit logging.

1.8 Chapter Summary

This chapter has presented the background to using Generative AI to detect social-media misinformation during emergencies. It has set out the challenge of whether LLM-based systems can distinguish reliable, false and unverified public-safety information, remain consistent and handle uncertainty appropriately. The chapter has also outlined the research aim, the three research objectives and the research questions that form the direction of the dissertation. The following chapters look at the relevant literature, research methodology, and experimental results and their implications in the context of using Generative AI responsibly for analysing emergency information.

Chapter 2: Literature Review

2.1 Overview of Literature Review

This literature review evaluates research on fake and deceptive public safety information in emergencies as well as the reliability of Generative AI in detecting such information. The review moves from the definition of information in emergencies and features of misinformation to their distribution via digital media. Detection methods are then discussed, from traditional approaches to artificial intelligence and Generative AI, together with reliability, accuracy, hallucination, context and evaluation. The recent literature indicates an increasing trend in automatic misinformation detection as well as the use of Generative AI. It also highlights persistent problems related to context, verification, consistency, and accuracy.

2.2 Public-Safety Information During Emergency Events

Public safety information refers to credible and actionable information that helps people or communities make informed decisions by understanding threats and taking necessary action. It comprises warnings, evacuation orders, health guidelines, information about shelters, information regarding changes in assessments of risks caused by natural disasters, weather extremes, pandemics, industrial accidents, terrorism and other public emergencies. The importance of this information goes beyond merely making it available because the accuracy and credibility of the information determine emergency decision-making. According to Badu *et al.* (2023), communication and trust play important roles in crisis management; therefore, information credibility plays an important role in emergency management. Emergencies are always urgent, and this makes the impact of incorrect information even more serious. As mentioned by Hu *et al.* (2024), the field of risk communication shows that information, trust and perceived ability to cope can affect individuals' preparedness decisions. Communication is connected to the protective behaviour of people. In this way, incorrect information can distort people's perception of risk and lead them to wrong actions.

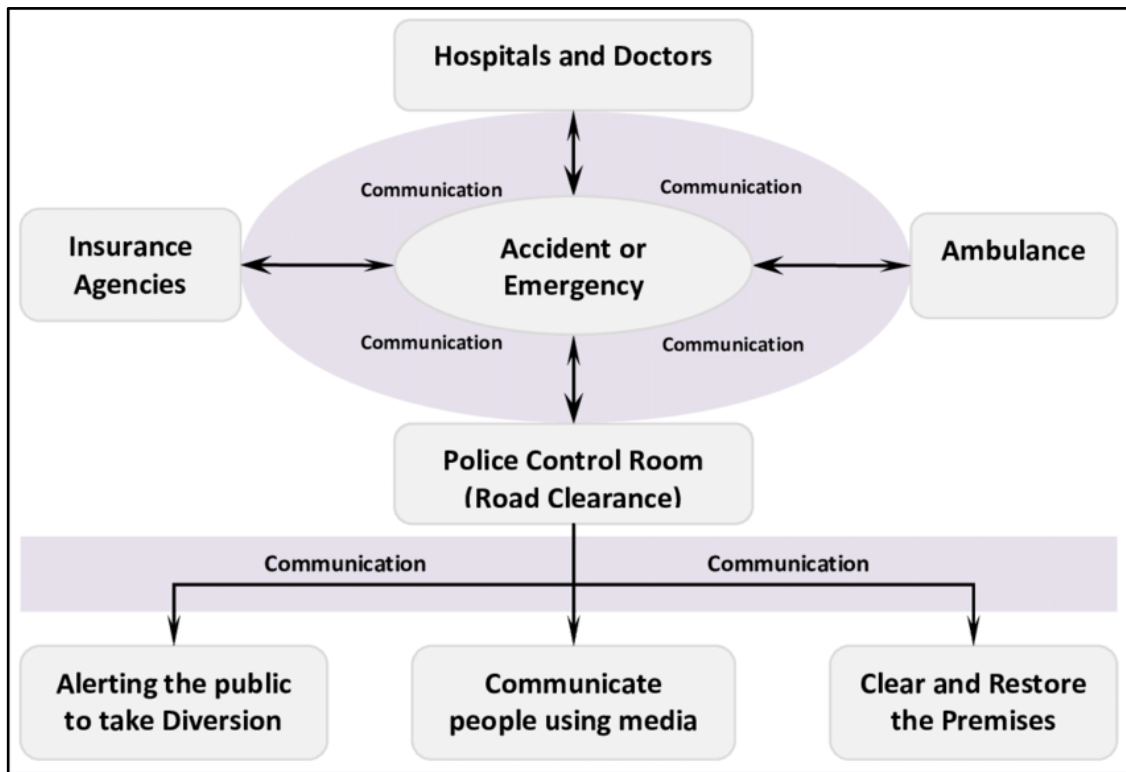


Figure 2.2.1: Performing important roles during emergency

(Source: Jadi, 2016)

The communication environment is a source of another crucial problem of reliability as well. Stolerio et al. (2024), studying emergency communication in seven countries, found that emergency authorities and the general public agree that social media is an essential mode of fast communication. Although social media offers advantages for emergency communication, there are issues relating to the reliability of information and its uncontrolled spread. Similarly, as argued by Stewart *et al.* (2023), dissemination, dialogue, and participation are the key components of disaster risk communication. The reliability of information is more of a public safety need than just a matter of preference for communication. Reliable information supports appropriate behavioural responses, builds confidence and improves preparation for emergencies, while false and misleading information will hamper decision-making and emergency management processes. This is why it is necessary to examine whether Generative AI can make such distinctions.

2.3 False, Misleading and Disinformation in Emergencies

False and misleading information is one of the major obstacles to public safety due to the fact that decisions made during emergencies are based on the availability of reliable information. It can be noted in the literature that various types of information that pose problems in terms of reliability are differentiated in terms of facts, intentions, degree of manipulation, and context.

El Mikati et al. (2023) show how these terms are distinguished in the literature. The differentiation is especially important for the current study, which goes further than deliberate lies and includes potentially false or misleading information.

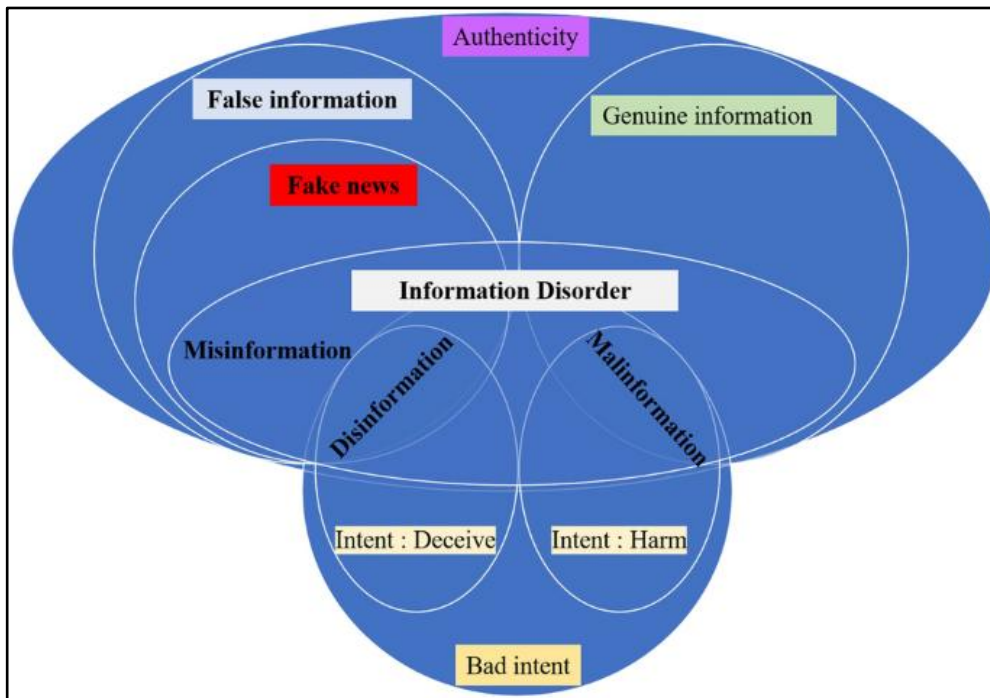


Figure 2.3.1: Fake news, disinformation and misinformation in social media

(Source: Aïmeur et al., 2023)

Misinformation, Disinformation and Malinformation

Misinformation is a term that is usually employed when referring to false information that is disseminated unintentionally. An example is an individual who shares misleading information about an emergency without realising it. The term disinformation describes the deliberate act of creating and sharing false information for the purpose of deceiving people. Malinformation is a distinct kind of information problem, because it involves accurate information used to cause harm. These distinctions show that detecting problematic information requires attention to intention as well as content, and not simply to truthfulness (Wardle and AbdAllah, 2023). These distinctions are especially important for assessing AI because an algorithm might recognise a lie while being unable to tell whether the message has been intentionally or accidentally made up.

False, Misleading, Rumour and Manipulated Content

This broader category includes misinformation, disinformation, rumours, fabrication and manipulation. False information does not match the facts, while misleading information results from selective information, distorted statistics, missing context or wrong interpretation. The fabrication of content implies the lack of factual basis, while manipulation implies the

transformation of truthful text, images, sound or video into misleading ones. Tomassi et al. (2024) show that the current information disorder is associated with the combination of automation, manipulation and misleading content.

Conceptual Definition for Generative AI Evaluation

A clear definition is needed in order to be able to evaluate Generative AI objectively. As Tsang (2026) notes, there are ongoing problems related to the ambiguity of definitions associated with fake information, in which the use of generic terms such as “fake news” may fail to capture distinctions among various kinds of information.

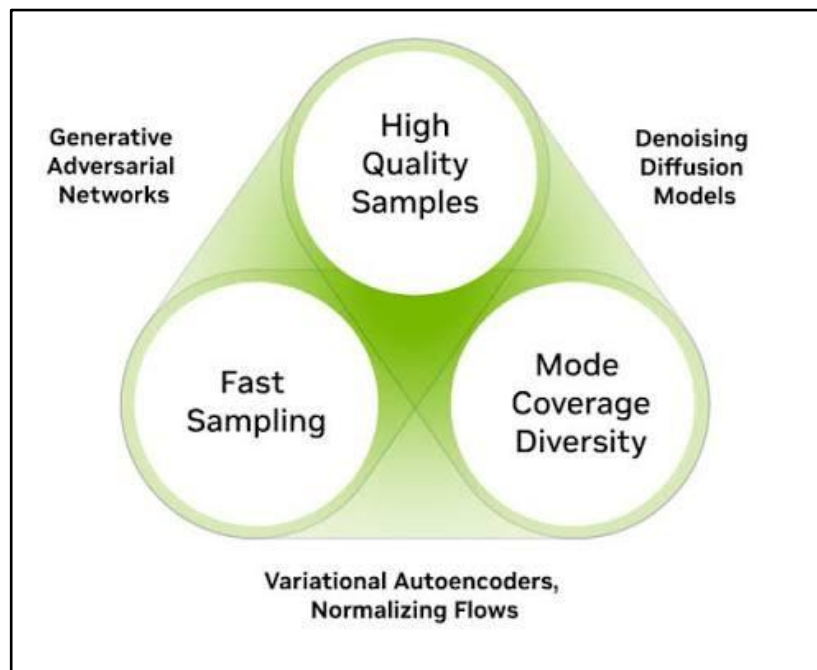


Figure 2.3.2: Generative AI Models

(Source: NVIDIA, n.d.)

In this study, false or misleading public-safety information means inaccurate or misleading information about public safety that is likely to distort perceptions of risk or of the appropriate action. This definition applies to any kind of misleading information regardless of the intention to deceive, which forms a standard for testing the accuracy of Generative AI.

2.4 Emergency Events and the Spread of Misinformation

Emergency-specific misinformation is different from regular misinformation as it arises in an unstable, information-rich and time-sensitive context. Natural calamities, pandemics, terrorist activities and severe weather patterns can produce partial information and shifting circumstances. These situations provide avenues for rumour-mongering, misinterpretation and deception to be spread even before reliable sources are able to authenticate them. Zhai *et al.*

(2024) highlight several key features that play an influential role in emergency-specific misinformation, which include uncertainty, information availability, emotion, time and context proximity.

Rapid Change, Uncertainty and Incomplete Information

The most significant aspect of emergency misinformation is that it relies heavily on the dynamic nature of information. In the course of an event, the facts related to deaths, evacuation areas, damaged infrastructure, climate/weather and aid available can change every minute or hour. A statement can therefore become incorrect even though no deception was intended. Lu et al. (2024) indicate the role played by uncertainty in the broader context of the information environment as a crucial element in the detection of misinformation, which includes uncertainty in relation to the physical environment, the media environment, communication networks, and message frames. Yang et al. (2023) show that disaster misinformation events need analysis in terms of their time and network features.

Table 2.4.1: Key Characteristics of Emergency-Related Misinformation

Characteristic	Emergency Context	AI Challenge	Reliability Concern
Rapid change	Facts change quickly	Assessing recency	Outdated claims
Uncertainty	Evidence may be incomplete	Ambiguous classification	False judgements
Emotional content	Fear and urgency increase sharing	Separating emotion from facts	Higher misinformation exposure
Location-specific claims	Information depends on place	Understanding context	Incorrect local decisions
Conflicting reports	Sources may disagree	Comparing evidence	Inconsistent outputs
Emerging events	Verification may be unavailable	Limited current evidence	Reduced accuracy

Conflicting Reports and the Reliability Challenge

An additional feature is conflicting information and the fast development of narratives. The emergence of information can occur from official organisations, journalists, witnesses, NGOs, and also regular social media users all at once during emergency communications. According

to Zhang et al. (2023), emergency information themes change over time as the event develops. This is a challenging detection environment because the lack of confirmation does not mean that something is false. This means that Generative AI requires more than simply a "true/false" system. Reliable evaluation requires considerations of recency, uncertainty, source context, location, conflicting evidence, and evolving information.

2.5 Social Media and Digital Platforms as Information Channels

Social media platforms have become major channels for disseminating information about emergencies (Li and Sheng, 2024). Users are capable of generating useful information regarding situations, including rescue appeals, damage reports, and updates on the ground. Shen et al. (2023), writing on disaster communication, confirm that social media can be used effectively in emergencies because of the exchange of information.

Speed, User-Generated Content and Network Effects

The speed and volume of digital communication are among its main strengths during emergencies. Information can be created and shared at a very fast pace such that vital messages for warnings and rescues can be transmitted to many people. Zhang et al. (2023) show through their study that the themes of emergency information change dynamically during the event cycle. Hocevar (2024) goes on to highlight clusters of communication networks during a wildfire emergency where citizen accounts play a vital part in information diffusion.

Misinformation Amplification and Digital Risks

Although social media facilitates communication during emergencies, the same features that make it helpful can also spread misinformation rapidly. This is because misinformation can spread through re-posting and from influential accounts and users connected through a network even before any validation is done. According to Zhai *et al.* (2024), social network structure, geographical proximity, and sentiment affect both the spreading of misinformation during disasters as well as correcting such misinformation.

Implications for Generative AI Reliability

The digital domain therefore leads to a dual effect on communication in that whereas platforms may use the quick availability of information for a good emergency response, they may at the same time lead to the increased availability of false or misleading information. A major challenge for Generative AI is that the information may come from various sources with varying levels of credibility, and through various channels and in different contexts.

2.6 Generative AI and Large Language Models for Classification

As discussed by Skumanich et al. (2025), large volumes of text data can be analysed and categorised in novel ways thanks to Generative Artificial Intelligence (AI) and the Large Language Models (LLMs) it has introduced. Typical classification systems will usually need specific datasets, feature engineering and machine-learning models which are trained on the task at hand. By contrast, LLM-based classification can accomplish these tasks with zero or few examples, using natural-language instructions. The broad capabilities of Generative AI could therefore be useful in analysing emergency-related social media data. Generative AI is most widely recognised for its capability to rapidly analyse large quantities of unstructured textual data, which is its primary link to information analysis in emergencies.

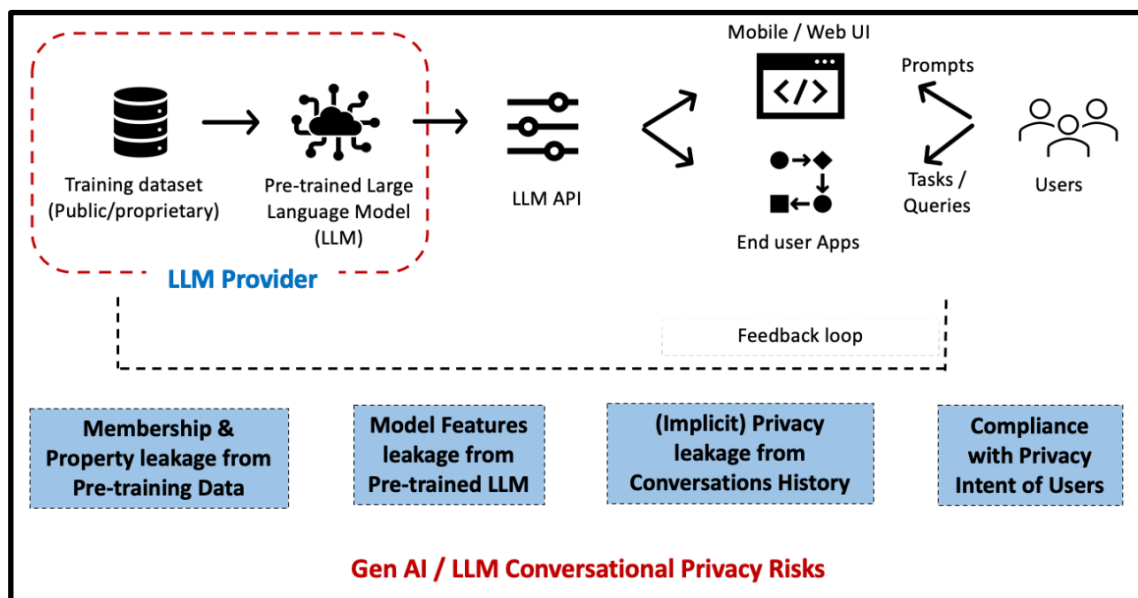


Figure 2.6.1: Generative AI and LLM-Based Information Classification Process

(Source: Vicari et al., 2026)

However, the ability of an LLM to classify information does not automatically establish that its decisions are reliable. Even when it lacks the knowledge needed, Generative AI can still provide fluent and convincing answers. There is therefore an inherent difference between a model's ability to classify and the reliability of its classifications. A model may associate linguistic patterns with misinformation without adequate evidence to determine whether a specific emergency claim is true. This is a crucial problem, particularly because of the contextual nature of information in emergencies. A statement may be true in one place but not in another, or may refer to a situation that has since changed. Authoritative evidence may also at times be unavailable, and the information may seem unverified at first (Bajwa, 2025). Unlike traditional classifiers, whose behaviour is fixed once they are trained, LLM classifications can vary with the instructions given. Even small changes in wording, order or complexity may

affect how the model interprets the task.

2.7 Reliability, Accuracy and Consistency of Generative AI

As examined by Ibrahim et al. (2026), reliability is more complicated than the mere classification accuracy. While accuracy gives an important clue as to how often the model makes the right classification, it does not show whether it is consistent or whether it makes errors in certain safety-relevant classification categories. A model can perform well overall and yet achieve low accuracy on a small minority class. This is especially crucial for public-safety applications, where a category may be rare but have significant consequences. Performance should therefore be assessed with more than one measure.

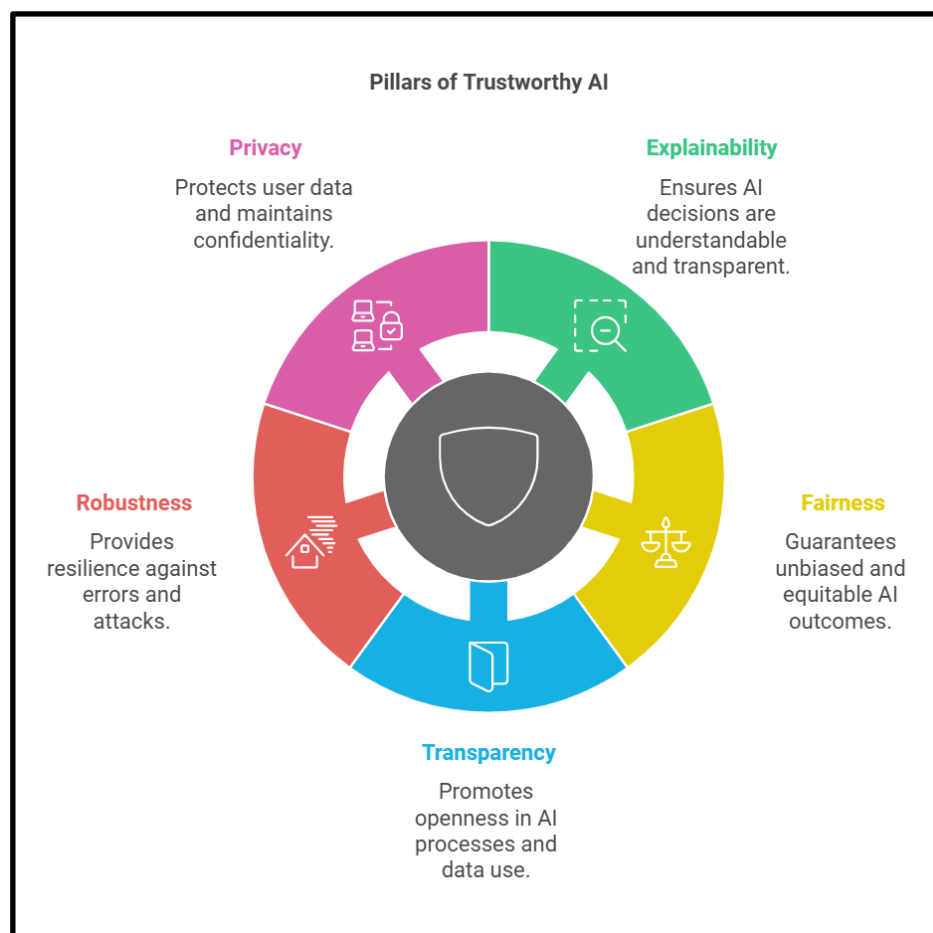


Figure 2.7.1: Dimensions of Generative AI Reliability

(Source: Shoaib et al., 2023)

Precision and recall provide two further views of classification behaviour. Precision measures how often predictions assigned to a category are correct, and recall measures the proportion of cases in a category that the system correctly identifies. Macro-F1 is also valuable because it provides a balanced assessment across classes. Specific patterns of misclassification can also be identified using confusion matrices. True warnings may be ignored if they are wrongly

flagged as false or misleading, while false information wrongly labelled as reliable gains unwarranted credibility. These errors therefore have different practical consequences and should not simply be combined into one overall error measure (Halford, 2026). When the same information is classified more than once under similar conditions and the model gives different answers, this indicates instability. Such behaviour can reduce trust in the model, even when its average accuracy seems high.

Table 2.7.1: Dimensions of Generative AI Reliability

Dimension	What It Measures	Importance to Emergency Information
Accuracy	Overall proportion of correct classifications	Provides an overall indication of performance
Macro-F1	Balanced performance across classification categories	Prevents common classes from dominating evaluation
Precision	Correctness of classifications assigned to a category	Identifies potentially incorrect classifications
Recall	Ability to identify cases belonging to a category	Important for detecting relevant misinformation
False Positives	Incorrect rejection or classification of reliable information	May cause genuine warnings to be disregarded
False Negatives	False information incorrectly treated as reliable	May increase the credibility of harmful information
Consistency	Stability across repeated evaluations	Indicates whether decisions are dependable
Prompt Sensitivity	Changes caused by different prompt formulations	Shows whether wording influences classification
Uncertainty	Ability to recognise insufficient evidence	Reduces inappropriate forced decisions

2.8 Hallucination, Unsupported Explanations and AI-Generated Errors

As demonstrated by Al-Shamrani and Hasan (2025), concerning the use of Generative AI in information verification, hallucination is a significant problem to address. This poses special potential hazards when the system is called upon to decide on the authenticity of information or whether it is misinformation during an emergency. Hallucination can make both the class and the explanation given by the model wrong. A model may produce a wrong classification, and it may also support a wrong classification with a seemingly convincing justification. The latter is especially problematic, because a detailed explanation may be taken as proof that the classification is correct.

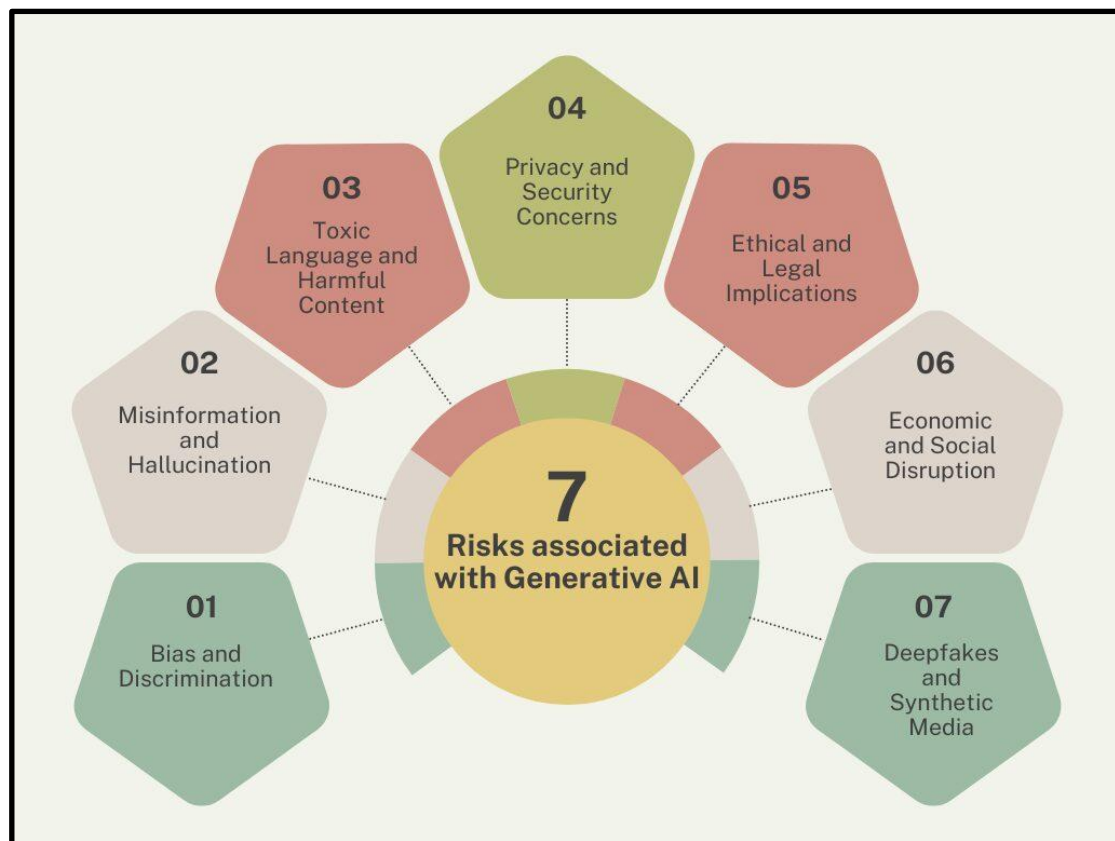


Figure 2.8.1: Hallucination and Unsupported Explanation Risks in Generative AI

(Source: Karuppanasamy et al., 2026)

An explanation is unsupported when it refers to evidence that does not exist. It is therefore not enough to look only at the final classification to judge whether the AI is reliable. Forcing the model to place every claim in one of the fixed categories can also encourage a decision even when there is not enough evidence. This can be an issue during emergencies as information may not be finalised. The model should be allowed to recognise information as unverifiable when there is not enough evidence, rather than inventing evidence or choosing between definite categories (Wang et al., 2026). The research must take into account the accuracy of the classification, support for the explanation, and uncertainty in the model. This matters because

a model might reach a correct answer for an incorrect reason, or give an incorrect answer with a very convincing explanation.

2.9 Human Verification and Human-in-the-Loop Decision Support

As highlighted by Kashyap Pargi *et al.* (2025), the constraints of Generative AI provide good reason for human validation. While AI systems can handle vast amounts of information quickly, decisions about misinformation in emergencies can affect the outcome of a crisis. Good judgment is always needed, especially in cases of uncertain, questionable or high-risk information. In a human-in-the-loop (HITL) approach, Generative AI handles the initial stages and human analysts review the important cases. It offers a compromise between the advantages of automated processing and the decision-making needed for the verification of complex information. This approach is particularly suitable for Unverified information. When a final decision is in doubt, the uncertain information can be passed to a human analyst for further investigation.

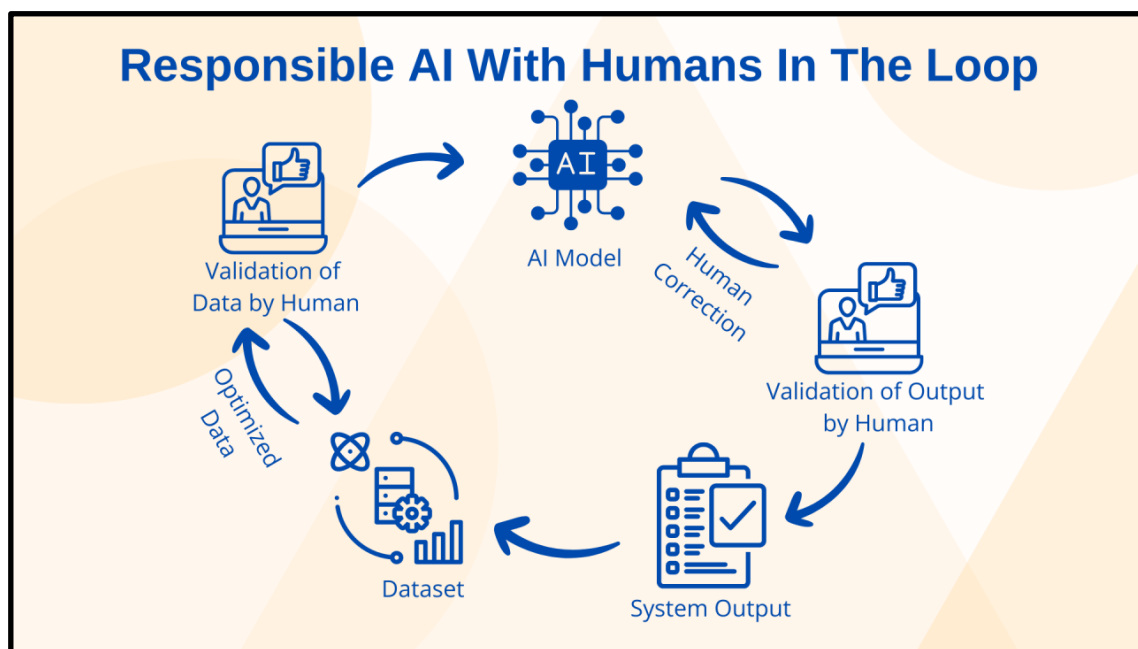


Figure 2.9.1: Human-in-the-Loop Verification for Emergency Information

(Source: Anggraini et al., 2026)

Human verification is also crucial when results are inconsistent. When the same claim is classified in two different ways under similar conditions, it can be a sign that the model's classification is "unstable". These cases should be examined, and not presumed to be reliable, on a case-by-case basis. There is also a risk of automation bias, in which analysts over-rely on AI outputs. A confident or detailed AI answer can seem more convincing than it deserves. Therefore, human verification should be based on real independent judgment, rather than

simply accepting the AI recommendation. The most appropriate role for Generative AI therefore appears to be supporting decisions rather than making them (Elhosary et al., 2026). The system can help with information triage, initial identification and classification of materials that could be an issue, but humans should make the final judgements.

2.10 Critical Review of Existing Research

There are existing studies that offer valuable insights into the concepts of emergency information, misinformation and Generative AI. Crisis informatics shows how social media can provide rapid information during a crisis, and emergency datasets can be used to classify disaster information and humanitarian information based on it. Rumour-verification research is more directly geared towards testing the veracity of information on the basis of evidence and conversational context. A major strength of previous research is its recognition of the benefits of automated systems in managing tremendous amounts of information created on social media that is generated in the event of an emergency, as well as the rate at which that information is created. The relevance of the emergency and its truth, however, are two different concepts (Buoncompagni, 2026). A post may be very relevant to an emergency, and yet it may be falsified, misleading or unverified. Thus, emergency classification does not necessarily reflect the result of reliable misinformation detection. Another drawback is the focus on performance metrics like accuracy, precision, recall and F1.

Table 2.10.1: Critical Review of Existing Research

Research Area	Main Strength	Key Limitation	Relevance
Crisis informatics	Emergency information analysis	Does not establish truthfulness	Emergency context
Social-media datasets	Realistic crisis content	Limited veracity focus	Emergency data
Rumour verification	Evidence-based veracity	Limited LLM reliability focus	Misinformation
LLM classification	Flexible classification	Reliability not guaranteed	AI approach

Prompt sensitivity	Identifies prompt effects	Limited emergency testing	Prompt evaluation
Hallucination	Identifies unsupported outputs	Limited integrated evaluation	Error analysis
Human-in-the-loop	Provides oversight	Escalation criteria needed	High-risk verification

2.11 Research Gap

The research gap is the lack of a comprehensive assessment of Generative AI's ability to detect false and misleading information in public-safety scenarios. Previous work has examined emergency information, rumour detection and LLM classification, but largely in isolation. A major shortcoming is that emergency relevance is often treated as though it indicated truthfulness. Another gap lies between reliability and accuracy, since repeated testing is required to evaluate whether AI decisions are consistent. It is also important to consider prompt sensitivity because there may be variation in classifications based on the wording of the prompts. Furthermore, traditional evaluations may fail to detect hallucinations or explanations that are not supported by evidence. Finally, emergency information can be incomplete, which makes uncertainty recognition and human verification necessary. The dissertation addresses these gaps through a multi-dimensional analysis of performance, consistency, prompt sensitivity, hallucination and uncertainty.

2.12 Conceptual Framework

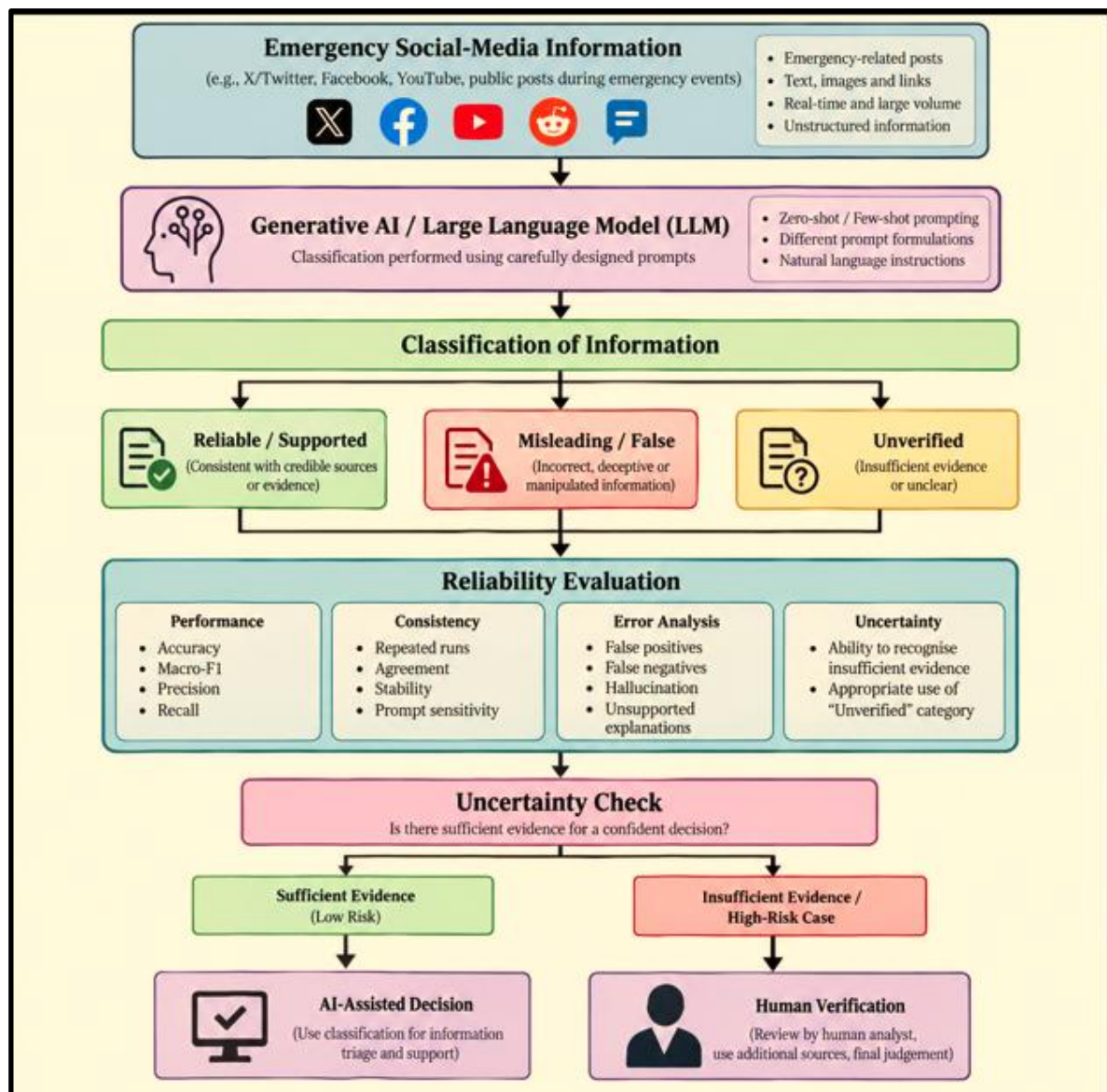


Figure 2.12.1: Conceptual Framework

2.13 Chapter Summary

This chapter critically reviewed the literature on Generative AI, Large Language Models, reliability, misinformation detection, hallucination, uncertainty and human verification in emergency information analysis. According to the review, Generative AI can efficiently classify vast amounts of emergency-related data quickly with flexible techniques. However, consistency, prompt stability and trustworthy explanations are required in addition to accuracy. Hallucinations and outputs that are not backed by facts could undermine assurance, especially for high-risk situations. In the case of ambiguous, uncertain and high-risk situations, human verification is thus critical. The chapter identified the need to integrate assessments of accuracy, consistency, prompt sensitivity, hallucination, uncertainty and safety-related errors, which

provides the theoretical foundation for this dissertation.

Chapter 3: Methodology

3.1 Research Approach

The study uses quantitative, experimental data analysis, supported by a qualitative analysis of the model's errors. The quantitative component measures the performance and reliability of Generative AI for classification of emergency-related content. The three classification categories that are covered in the study are Reliable/Supported, Misleading/False and Unverified. These categories enable the research to distinguish information that is supported by evidence, contradicted by evidence, or lacking sufficient evidence. The approach focuses on reliability, not only accuracy. The chosen Large Language Model (LLM) will be evaluated based on various metrics including accuracy, Macro-F1, precision, recall, false positive and false negative rates, and confusion matrix (Mukherjee, 2026). Repeated testing will also be used to check whether classifications remain stable when the same items are classified several times.

Three prompting conditions—zero-shot, definition-based and few-shot prompting—will be investigated. Under these conditions, prompt variants that change synonyms or word order without changing the meaning will be tested. As a result of this, the research can evaluate the impact of seemingly incidental changes in instructions for classifications. The quantitative results will be supplemented by qualitative error analysis. Emphasis will be placed on hallucinations, unsupported explanations, forced classifications and misidentifications related to safety (Phillips and Chan, 2025). Together, these methods test whether Generative AI can give accurate, consistent and cautious responses to emergency public-safety information.

3.2 Research Design

The study employs a controlled experimental research design to test Generative AI in various scenarios and with diverse prompts. The process proposed is:

Dataset selection → Pre-processing → Prompt generation → LLM inference → Output parsing → Result storage → Statistical analysis → Error analysis.

This procedure offers a structured method to keep track of the original information, the prompt, model answer and final output. Additionally, it allows for planned comparisons across prompting conditions. The design initially considered HumAID or CrisisBench as sources of emergency-relevance labels and PHEMEPlus or RumourEval as sources of veracity labels; the final implementation uses PHEMEPlus, because it provides veracity annotations. Emergency-relevance labels will not automatically be treated as truth labels (Chen, 2025). Instead, the study

will be based on evidence or on existing ground truth information in order to identify Reliable/Supported, Misleading/False, and Unverified cases. Ambiguous items will not be forcibly placed into one of the categories.

Three conditions are tested: zero-shot prompting (classification using the label names only), definition-based prompting (classification using explicit definitions of each label), and few-shot prompting (classification with a small number of labelled examples). Within each condition, several semantically equivalent variants of the prompts will be tested. The chosen sample will also be tested on multiple occasions in a controlled environment. The model identifier, experiment date, full prompt, configuration parameters, input information, unedited response, classification and explanation will be recorded (Zhang et al., 2026). Configuration variables are held constant across comparisons and are changed only during sensitivity testing. The design therefore keeps conditions controlled and is easy to reproduce.

3.3 Research Strategy

This dissertation adopts a controlled experimental strategy that combines quantitative assessment with qualitative error analysis. This directly reflects the research aim of evaluating how reliably Generative AI handles false, misleading and supported public-safety information during crises. The approach does not consider the relevance of emergencies to be the same as truthfulness. Rather, it determines how well a Generative AI model can differentiate between Reliable/Supported, Misleading/False and Unverified information through a repeated experimental procedure.

The underlying principle is that reliability goes beyond achieving a high proportion of accurate classifications. The literature review showed that an LLM can give fluent and convincing answers while still giving incorrect classifications, unsubstantiated answers or overconfident decisions (Al-Shamrani and Hasan, 2025). As a result, performance, consistency, prompt sensitivity, uncertainty and error-related dimensions are built into the experimental strategy. The approach aligns with the conceptual framework identified in Chapter 2, where Generative AI classification is followed by reliability testing and a check on uncertainty.

The PHEMEPlus rumour dataset is the main source of emergency-related social-media data. The final implementation obtained 1,972 rumour records, each containing the event, thread ID, tweet ID, source text, claim, PHEME label, research label and annotation information. The dataset covers five emergency events: Charlie Hebdo, Ferguson, Germanwings, the Ottawa shooting and the Sydney siege. These events provide a range of emergency situations and

information conditions, allowing a wider evaluation of model behaviour than a single event would (Altayeva, 2025).

The strategy also includes a balanced sampling procedure. The complete dataset contains 1,008 Reliable/Supported records, 393 Misleading/False records and 571 Unverified records. Owing to the uneven distribution of these categories, 100 records were sampled from each category, giving a balanced experimental sample of 300. This reduces the chance that the largest category dominates the assessment and allows the three classes to be compared more directly.

The strategy also controls the prompting environment. There are three prompting conditions, Zero-Shot, Definition-Based and Few-Shot prompting. Under these circumstances, semantically equivalent prompt versions will be used to test the effect of wording change on what the model decides. The repetition of the same information also enables the research to differentiate between actual performance and some accidental or volatile responses. This directly addresses the research questions on prompt sensitivity and consistency.

3.4 Research Method

The research adopts a quantitative experimental research design which is backed by the qualitative analysis of errors. Generative AI classification performance and reliability are assessed in the quantitative component whereas in the qualitative component incorrect, unsupported and potentially unsafe responses are studied.

The model that is chosen is openai/gpt-oss-120b. The implementation also recorded an “API TEST: SUCCESS” message, confirming that the model connection worked before the experiment.

Table 3.4.1: Operational Definitions of the Research Categories

Research category	Operational meaning
Reliable/Supported	Information supported by evidence or ground-truth information
Misleading/False	Information contradicted by evidence or identified as false/misleading
Unverified	Information without sufficient evidence for a reliable or false classification

The experiment uses Zero-Shot, Definition-Based and Few-Shot prompting. Outputs of all three approaches can be found in the completed pilot. Each condition is designed to be run

three times to test consistency, and each category contributes 100 observations, giving a balanced experimental sample of 300 records.

Qualitative analysis of false positives, false negatives, hallucinations, unsupported explanations, forced classifications and safety-related misidentifications is also part of the approach. This makes sure that the evaluation is not just based on the overall accuracy but also the nature and the potential consequences of model errors (Bajwa, 2025).

3.5 Data Collection and Analysis

The data come mainly from the completed extraction and preparation of the PHEMEPlus dataset. The implementation recognised five directories of emergency-events and was able to extract the rumour information. The final dataset includes 1,972 records, with no blank source-text values. The number of records that have evidence URLs is 1,848 and that without an evidence URL is 124, which shows that evidence is not available evenly across the dataset.

The research-label mapping produced 1,008 Reliable/Supported records, 393 Misleading/False records and 571 Unverified records. Every rumour annotation was mapped to one of the three research categories, and the final output reports no unmapped cases.

The main comparative experimental data are based on a balanced sample of 300 records. Each research category has 100 observations. With this method of sampling, the three classes receive equal representation although there was imbalance in the original PHEMEPlus data. A parallel comparison of the performance at the class level can thus be conducted more effectively using the sample.

The Python workflow handles data extraction, data preparation, prompt construction, model prediction, result formatting and response storage. The results of the model answers are compared to the defined research labels. Accuracy, precision, recall, Macro-F1, false-positive rate, false-negative rate and confusion matrices are the main quantitative measures. These measures reflect the multidimensional definition of reliability set out in Chapter 2.

Table 3.5.1: Evaluation Measures and Their Purpose

Measure	Purpose in the study
Accuracy	Measures the overall proportion of correct classifications
Precision	Measures the correctness of predictions assigned to each category

Recall	Measures the ability to identify cases belonging to each category
Macro-F1	Provides balanced evaluation across the three classes
False positives	Identifies reliable information incorrectly treated as problematic
False negatives	Identifies misleading/false information incorrectly treated as reliable
Confusion matrix	Shows detailed patterns of classification errors
Consistency	Examines agreement across repeated evaluations
Prompt sensitivity	Examines changes associated with semantically equivalent prompts

The results distribution across emergency events is also taken into account in the analysis. This matters because the final dataset varies considerably from event to event. To illustrate this, there are 522 threads in Sydney, 470 in Ottawa and 458 in Charlie Hebdo, and 284 and 238 in Ferguson and Germanwings respectively. This variability will give valuable background to the classification performance.

Qualitative analysis follows up the numerical results by examining examples of faulty or wrong outputs. Special attention is paid to hallucinated explanations, references to evidence that does not exist, and forced classifications. This follows the literature review, which notes that a model may generate a plausible explanation that lacks factual support (Chen, 2025). The analysis then takes into consideration not only which classification the model would give, but also why the response can be a problem.

The pilot also indicates that the model was able to generate responses in all three prompting conditions with many responses returned in less than one second. The outputs included Unverified, Reliable/Supported and Misleading/False classifications, as well as some unsuccessful API responses. These observations show that both response quality and technical outcomes should be considered in the analysis.

3.6 Validity and Reliability

Validity and reliability are important because the research tests the stability of Generative AI when sorting information about emergencies. Separating emergency relevance from information veracity supports construct validity. Reliable/Supported, Misleading/False and

Unverified are research categories guided by evidence and ground-truth information instead of presupposing that all information about emergencies will be correct.

Internal validity is supported by controlled experimental conditions, consistent classification categories and systematic recording of model inputs, prompts, responses and classifications (Chowdhury, 2026). Three repeated evaluations, built into the experimental design, permit the assessment of consistency in repeated classifications.

The balanced sample of 300 observations (100 records per research category) further supports reliable comparison. The dataset consists of 1,972 records with unique rumour and tweet identifiers, which supports data consistency. Evidence-based evaluation is also supported by the fact that 1,848 records have evidence URLs, whereas 124 records without these URLs confirm that the Unverified category is worth retaining instead of forcing those records into the Reliable/Supported or Misleading/False categories.

3.7 Ethical Consideration

Ethical considerations are also crucial because the study deals with misinformation and public-safety information during emergencies. The study draws on available secondary data on social-media and rumours, as opposed to gathering personal data on the participants (Guo et al., 2023). The methodology is thus not about interviews, questions or personal communication with those individuals, whose posts are represented in the dataset.

Analysis is made of claims, classifications and model behaviour, instead of trying to determine or profile the individual social-media users. User identifiers included in the original data set are dealt with as dataset fields needed to accomplish record linkage and assessment and not as an object of individual analysis.

The other ethical concern is the possible results of false AI classifications. The dissertation recognises that an incorrect classification of emergency information can have serious consequences. False negatives can make damaging misinformation seem truthful, whereas false positives can cause genuine warnings or precautions to be dismissed. Because of this, the study does not present Generative AI as an independent decision-maker in an emergency.

The Unverified category is also adopted as part of the methodology as a measure of ethical protection. In cases where there is a lack of evidence the information is not necessarily considered to be true or false. This minimises the chance of creating false certainty based on partial evidence. Likewise, qualitative analysis of hallucinatory and non-evidenced justifications pinpoints instances where a credible-looking model answer lacks factual support.

The research also supports a human-in-the-loop principle for uncertain and high-risk cases. The review of the literature confirms that Generative AI is better suited as decision support and not as a decision-maker. This is especially applicable to emergency information, where human judgement is essential when evidence is incomplete or conflicting, or when safety is at stake. Lastly, the experimental procedure is transparent with documentation of model configuration, prompts, inputs and outputs. This kind of documentation can facilitate auditability and minimise the chances of only reporting successful classifications. Ethical assessment is thus a component of the greater reliability system rather than a separate administrative requirement.

3.8 Summary

In this chapter, the methodology of the reliability assessment of Generative AI to classify emergency public-safety information was laid out. The study takes a controlled experimental approach that combines quantitative performance measurement with qualitative error analysis. The final implementation is based on the PHEMEPlus dataset and comprises 1,972 extracted rumour records of five emergency events. The data were coded as Reliable/Supported, Misleading/False or Unverified, and a balanced sample of 300 records, 100 per category, was drawn.

The final experiment is based on the openai/gpt-oss-120b model with the implementation of Zero-Shot, Definition-Based and Few-Shot prompting. Consistency is evaluated through three repeated runs, and accuracy, precision, recall, Macro-F1, false-positive and false-negative rates and confusion matrices are used to assess multidimensional performance.

Evidence-based research labels, balanced sampling, controlled experimental conditions and recording of inputs and outputs are strategies which reinforce validity. Reliability is checked through repeated classifications and prompt-sensitivity testing. Ethical considerations are addressed through responsible handling of emergency information, avoiding forced classification, attention to hallucinations and unsubstantiated explanations, and recognition that human verification is essential in doubtful or hazardous cases.

In general, the methodology offers a direct connection between research questions, software implementation and evaluation presented in Chapter 4. It provides a sound basis for interpreting the Generative AI outcomes without treating model fluency or a single accuracy figure as sufficient evidence of reliability.

Chapter 4: Results and Analysis

4.1 Introduction

This chapter presents the outcomes of implementing and assessing Generative AI for classifying supported, false and misleading public-safety information during emergencies. The PHEMEPlus data analysed contain 1,972 rumours covering five emergency events: the Charlie Hebdo attack, Ferguson, Germanwings, the Ottawa shooting, and the Sydney siege. Records were sorted into three research labels: Reliable/Supported, Misleading/False and Unverified. Implementation was in line with the experimental procedure outlined in Chapter 3 including balanced sampling, controlled prompting and planned repeated evaluation.

The analysis is based on the quality and structure of the retrieved data, the proportion of each research label, the availability of evidence, the variability across the five emergency events, and API behaviour during pilot testing. The model selected was openai/gpt-oss-120b, and the main experimental settings are Zero-Shot, Definition-Based and Few-Shot prompting. Three repeated evaluations were planned to test consistency and prompt sensitivity. This chapter therefore assesses reliability beyond a single accuracy figure, addressing classification behaviour, experimental stability, API performance and the potential risks arising from AI classifications that are not supported or forced.

4.2 Implementation and Evaluation

```
print("Dataset:", DATASET_ROOT)
print("Model:", MODEL_NAME)
print("Provider:", HF_PROVIDER)
print("Repeats:", REPEATS)
print("Target per class:", TARGET_PER_CLASS)
```



```
Dataset: D:\PHEME_PLUS\all-rnr-annotated-threads
Model: openai/gpt-oss-120b
Provider: auto
Repeats: 3
Target per class: 100
```

Figure 4.2.1: Experimental Configuration

The experimental setup for the PHEMEPlus analysis is shown in Fig 4.2.1. The data are stored in the D:\PHEME_PLUS\all-rnr-annotated-threads folder and the configured model is

openai/gpt-oss-120b. To keep the evaluation structured and reproducible, three repeats are programmed, with 100 samples per research-label class (Park and Nan, 2026).

```
print("API TEST: SUCCESS")
print("Model:", MODEL_NAME)
print("Response:", repr(text))

except Exception as e:
    print("API TEST: FAILED")
    print("Error type:", type(e).__name__)
    print("Error:", str(e))
    raise
```

API TEST: SUCCESS
Model: openai/gpt-oss-120b
Response: 'API_TEST_SUCCESS'

Figure 4.2.2: API Test Success

Figure 4.2.2 shows that API connectivity was established before the main experiment. The output reads “API TEST: SUCCESS” and confirms that the model in use is openai/gpt-oss-120b. The successful response shows that the API connection worked at the test stage, which justified continuing with model-based classification.

```
event_dirs = sorted(p for p in DATASET_ROOT.iterdir() if p.is_dir())

print("Events found:")
for p in event_dirs:
    print(" -", p.name)

if not event_dirs:
    raise RuntimeError("No event directories found.")
```

Events found:

- charliehebdoo-all-rnr-threads
- ferguson-all-rnr-threads
- germanwings-crash-all-rnr-threads
- ottawashooting-all-rnr-threads
- sydneyseige-all-rnr-threads

Figure 4.2.3: Dataset Events

Figure 4.2.3 identifies the five emergency-event directories contained within the PHEMEPlus dataset. They are Charlie Hebdo, Ferguson, Germanwings, the Ottawa shooting and the Sydney siege. The output shows that the five event directories were correctly detected in the data, thus confirming the dataset structure needed for rumour extraction and analysis across various emergency contexts.

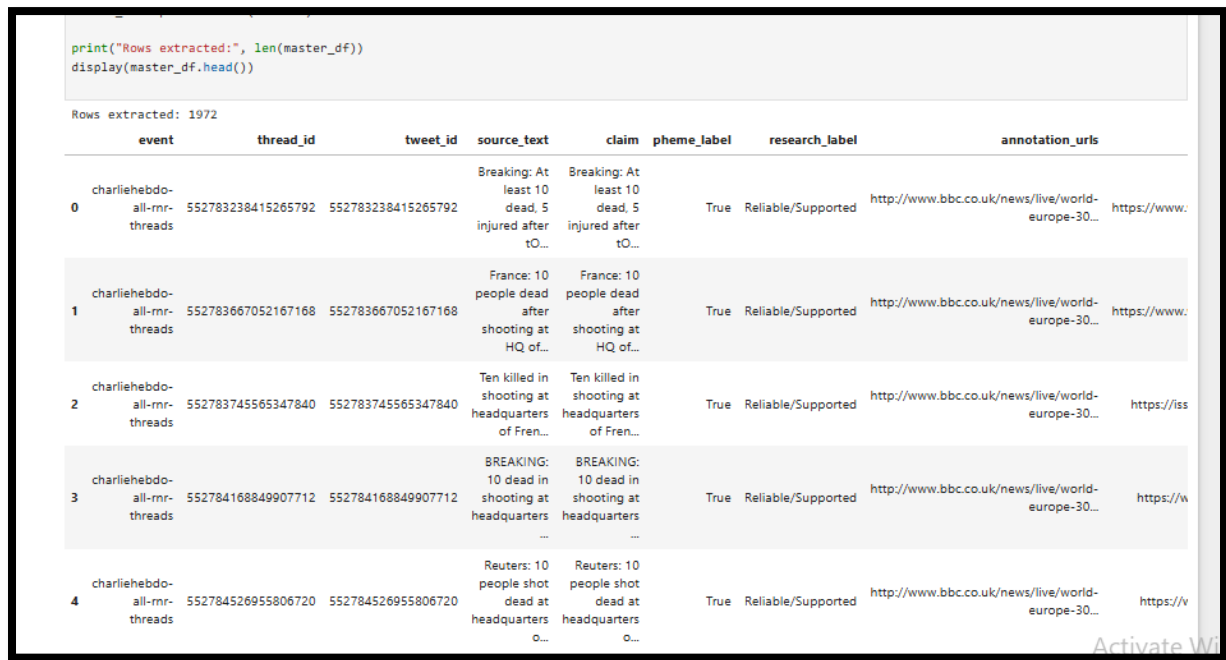


Figure 4.2.4: Extracted Rumour Dataset

The rumour dataset retrieved is shown in Figure 4.2.4 with 1,972 records. Each record contains the event, thread ID, tweet ID, source text, claim, PHEME label, a research label and annotation URLs. The example shows source tweets matched to their ground-truth classifications, which supports systematic evaluation of Generative AI (Kumar et al., 2025).

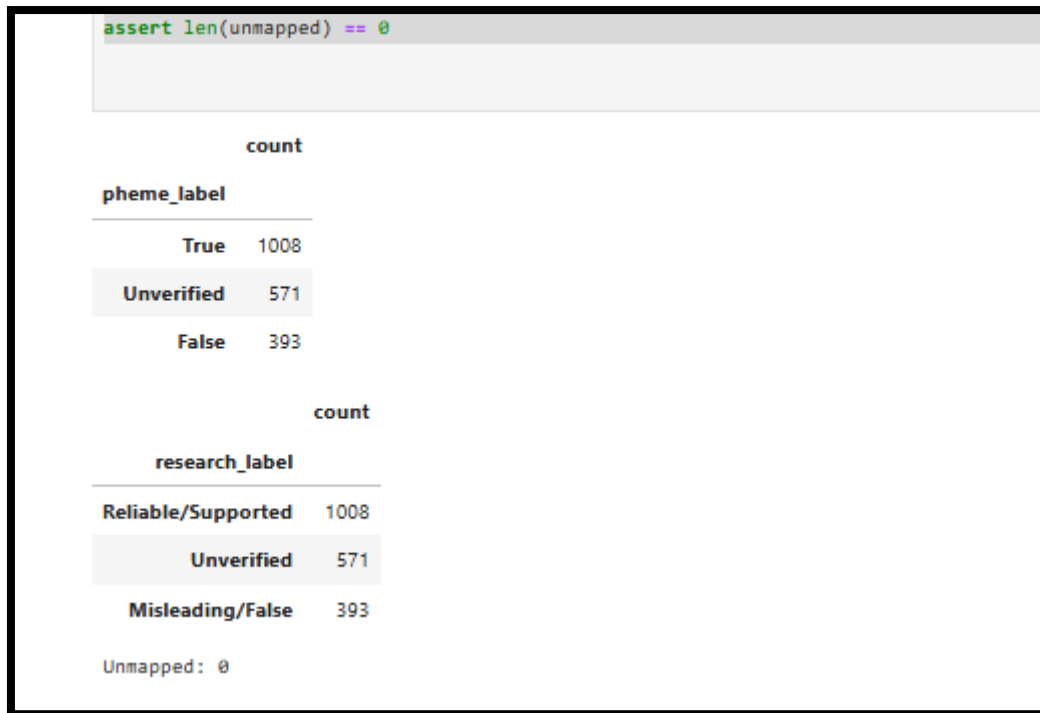


Figure 4.2.5: Ground-Truth Label Validation

The results of the research-label mapping are presented in Figure 4.2.5. There are 1,008 Reliable/Supported records, 393 Misleading/False records and 571 Unverified records. The output report shows that every extracted rumour annotation was assigned to one of the three research categories, with no unmapped cases.



Figure 4.2.6: Dataset Quality Check

The structural quality of the extracted data is shown in Figure 4.2.6. The dataset has 1,972 rumour threads, 1,972 distinct rumour IDs and 1,972 distinct tweet IDs. It is complete, with no empty source-text fields; 1,848 records have an evidence URL and 124 do not. The quality check confirms that the dataset is complete in terms of identifiers and source text, which minimises problems associated with missing observations. Whether a record has an evidence URL matters, because an evidence reference makes it easier to judge whether the record is supported, misleading or unverified. Both the completeness of the dataset and the availability of evidence therefore affect the interpretation of model reliability.

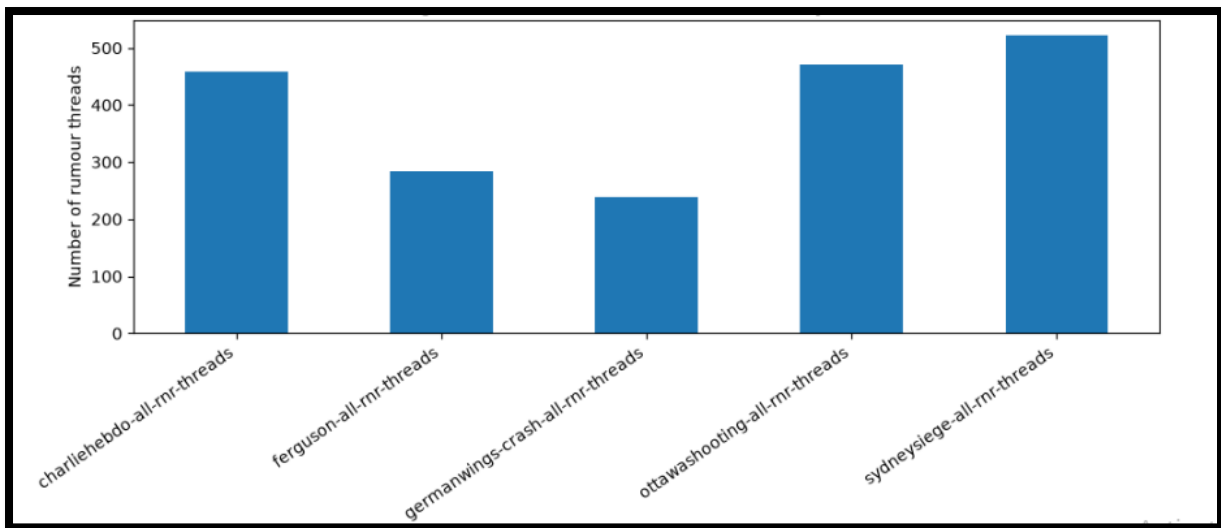


Figure 4.2.7: Rumour Threads by Event

Figure 4.2.7 shows how the 1,972 rumour threads are distributed across the five emergency events. Sydney has the most, with 522 threads, followed by the Ottawa shooting (470) and Charlie Hebdo (458), while Ferguson (284) and Germanwings (238) have fewer. Some events therefore contain many more rumours than others, suggesting that the PHEMEPlus dataset represents a range of social-media engagement during emergencies. This variation is important because event-specific language, circumstances and information characteristics can have an influence on model behaviour. The structure also allows the study to consider whether classification reliability depends on the emergency context.

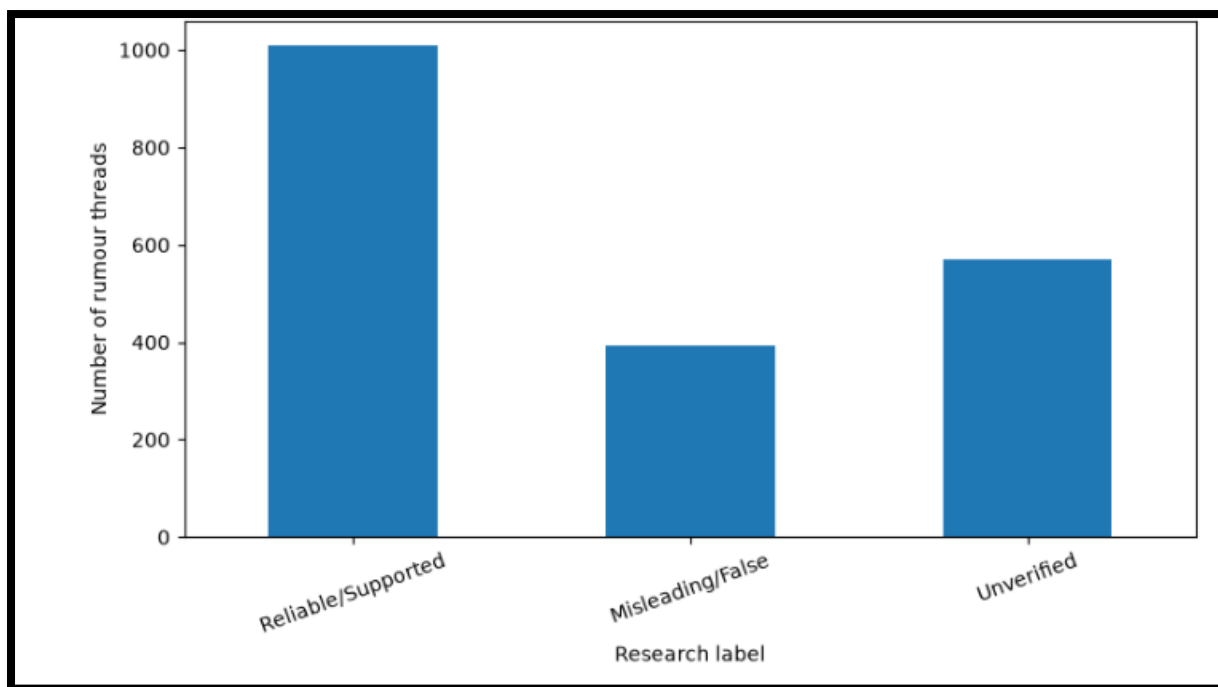


Figure 4.2.8: Ground-Truth Label Distribution

The overall distribution of the ground-truth research labels is shown in Figure 4.2.8. The highest number of threads is for Reliable/Supported at 1,008, followed by Unverified at 571 and Misleading/False at 393. The distribution shows that the complete data set is imbalanced, thus justifying the usage of a balanced experimental sample to evaluate the model. If classification performance were measured only on the original distribution, the Reliable/Supported class would dominate the results. Equal sampling instead gives each research label the same weight in the comparative evaluation, which helps to reveal Generative AI behaviour on supported, misleading and uncertain information.

	research_label		
	Reliable/Supported	Misleading/False	Unverified
event			
charliehebd0-all-rnr-threads	193	116	149
ferguson-all-rnr-threads	10	8	266
germanwings-crash-all-rnr-threads	94	111	33
ottawashooting-all-rnr-threads	329	72	69
sydneyseige-all-rnr-threads	382	86	54

Figure 4.2.9: Ground-Truth Veracity by Event

The ground-truth veracity counts are shown in Figure 4.2.9 for the five emergency events. There are 193 Reliable/Supported, 116 Misleading/False and 149 Unverified cases in Charlie Hebdo. The composition differs markedly between events, as the Ferguson figures show: 10 Reliable/Supported and 8 Misleading/False, but 266 Unverified cases. The event-level differences highlight that there is no uniform veracity distribution for emergencies. Ferguson is dominated by Unverified records, while Charlie Hebdo has a more even distribution. These differences matter because supported, uncertain and misleading information must be distinguished across whole emergency events, not just case by case.

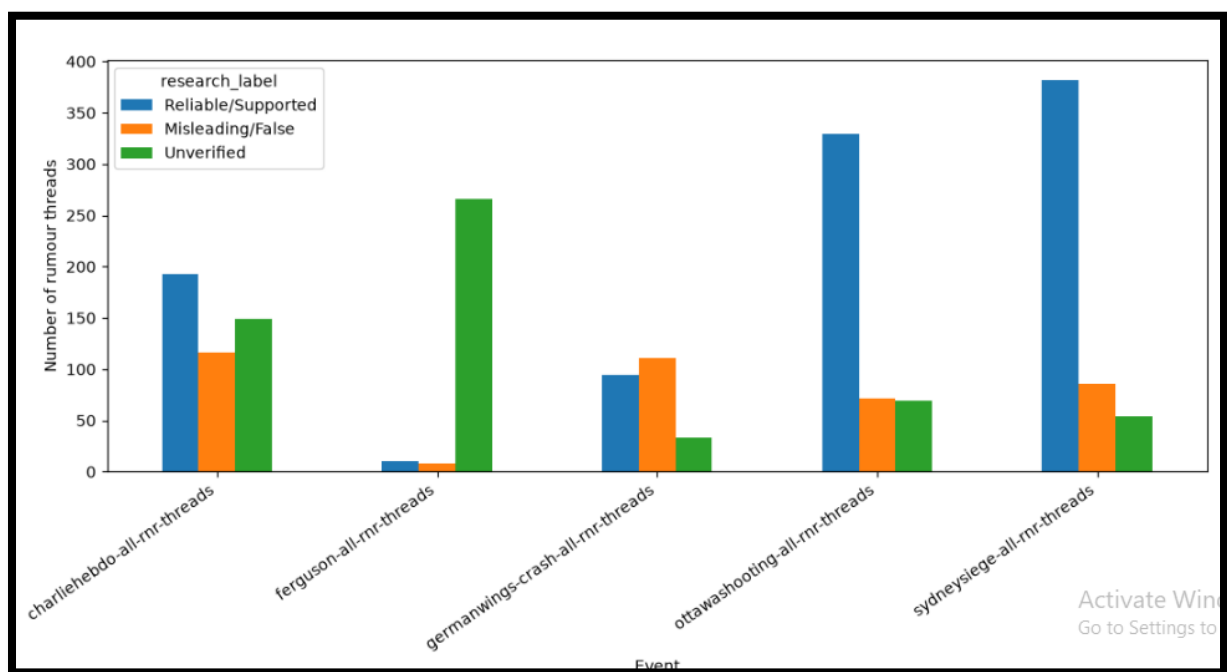


Figure 4.2.10: Distribution of Ground-Truth Veracity by Event

Figure 4.2.10 provides a picture of the ground truth veracity differences between emergency events. Sydney has the most Reliable/Supported cases (382), followed by the Ottawa shooting (329). Germanwings has 111 Misleading/False cases, while Ferguson has 266 Unverified cases. The figure shows, therefore, a high level of variability of events. The difference demonstrated by the events highlights the need for the assessment of Generative AI in different emergency situations. A model that performs well on one event's mix of labels may perform quite differently on an event dominated by Unverified claims. Event-level distributions therefore help to show where a model falls short, which aggregated measures may mask.

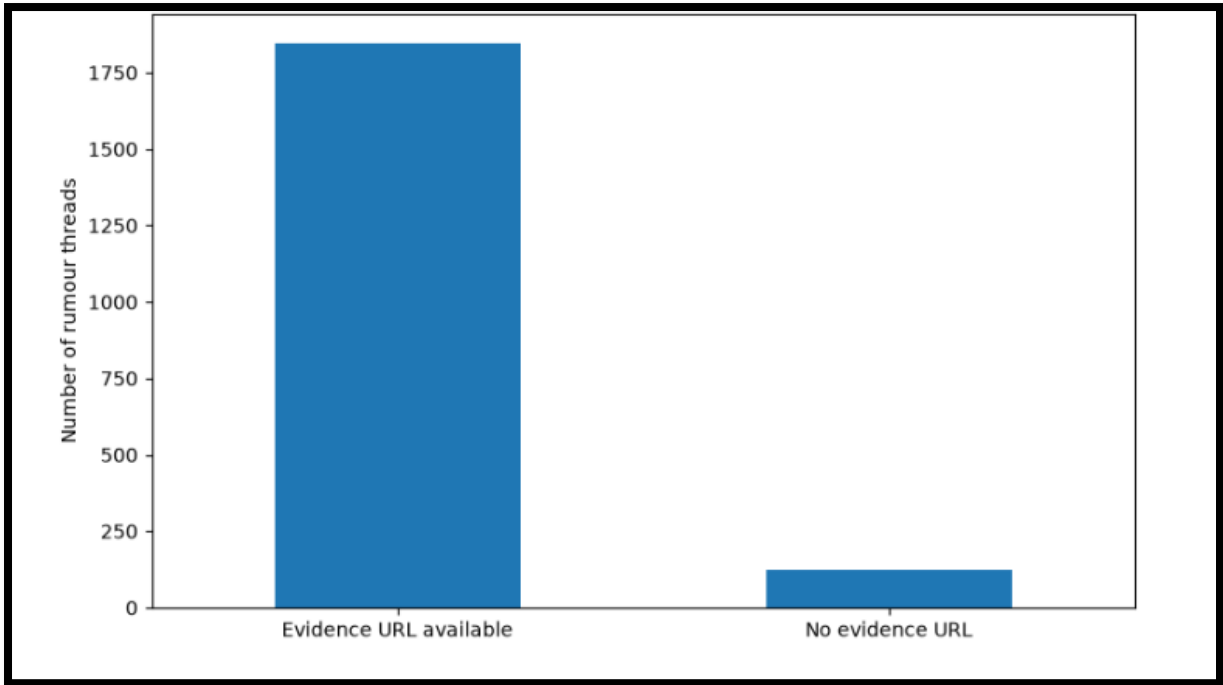


Figure 4.2.11: Evidence URL Availability

Figure 4.2.11 shows the evidence URL availability in PHEMEPlus rumour dataset. There are 124 threads where there is no evidence URL, and 1,848 threads with evidence URLs. Most records therefore have associated evidence references, while a smaller set has none. The reason for mentioning evidence availability is that in the study, uncertainty is recognised as a valid classification outcome when information is absent. The records with evidence URLs provide a basis for evidence-assisted analysis, while those without show cases in which verification may be more difficult. This distinction supports the research focus on careful classification that avoids overstated certainty.

available	
research_label	
Reliable/Supported	1008
Misleading/False	393
Unverified	571
Experiment sample: (300, 11)	
sample_count	
research_label	
Reliable/Supported	100
Misleading/False	100
Unverified	100

Figure 4.2.12: Experimental Sample Distribution

Figure 4.2.12 shows the balanced experimental sample drawn from the available dataset. The full dataset contains 1,008 Reliable/Supported, 393 Misleading/False and 571 Unverified records. One hundred observations are sampled from each class to create a balanced sample of 300 observations for comparative evaluation of Generative AI. The balanced sample allows direct comparisons of Generative AI performance across the three research categories. Each class contributes exactly 100 observations, so the original imbalance does not dictate the label proportions. This design is aligned with the goal of measuring classification reliability and helps to interpret performance differences directly across Reliable/Supported, Misleading/False and Unverified information.

```

[
  {
    "thread_id": "544512691948498944",
    "research_label": "Reliable/Supported",
    "source_text": "Gunfire at Sydney cafe where hostages held; re..."
  },
  {
    "thread_id": "500287224957898752",
    "research_label": "Misleading/False",
    "source_text": "Being shot 10 times in the street for stealing..."
  },
  {
    "thread_id": "553468492700454912",
    "research_label": "Unverified",
    "source_text": "BREAKING NEWS - Exchange of fire reported in #..."
  }
]

```

	thread_id	research_label	source_text
0	544512691948498944	Reliable/Supported	Gunfire at Sydney cafe where hostages held; re...
1	500287224957898752	Misleading/False	Being shot 10 times in the street for stealing...
2	553468492700454912	Unverified	BREAKING NEWS - Exchange of fire reported in #...

Figure 4.2.13: Experimental Sample Preview

Figure 4.2.13 is a preview of the experimental sample selected. The records shown include the thread ID, source text and research label for claims in each of the three categories. The preview shows that the final sample includes the content of the original social-media claims and is evenly distributed among the three categories of the research labels. This matters because classification reliability depends on analysing the real language of emergency social-media posts. Preserving the source text also makes model predictions traceable to ground-truth labels, which helps in understanding errors on specific claims.

```

Pilot: 100% [Progress Bar] 5/5 [00:57<00:00, 14.05s/it]

Attempt 1/5 failed: HfHubHTTPError: Client error '402 Payment Required' for url 'https://router.huggingface.co/v1/chat/completions' (Request ID: Root=1-6aa5308d-45e6347952b615a02224742;148f1842-ea88-4699-9cd5-e881b60f3f46)
For more information check: https://developer.mozilla.org/en-US/docs/Web/HTTP/Status/402

You have depleted your monthly included credits. Purchase pre-paid credits to continue using Inference Providers. Alternatively, subscribe to PRO to get 20x more included usage.
Waiting 1 seconds...

Attempt 2/5 failed: HfHubHTTPError: Client error '402 Payment Required' for url 'https://router.huggingface.co/v1/chat/completions' (Request ID: Root=1-6aa5308e-12f58d4e5226b6b133f9d62f;4247bb17-6344-41b2-8f03-48d788d71406)
For more information check: https://developer.mozilla.org/en-US/docs/Web/HTTP/Status/402

You have depleted your monthly included credits. Purchase pre-paid credits to continue using Inference Providers. Alternatively, subscribe to PRO to get 20x more included usage.
Waiting 2 seconds...

Attempt 3/5 failed: HfHubHTTPError: Client error '402 Payment Required' for url 'https://router.huggingface.co/v1/chat/completions' (Request ID: Root=1-6aa53090-5de6fc8c779c6df6221d9cca;f4d05160-2d84-4291-a79d-d6e8e9806964)
For more information check: https://developer.mozilla.org/en-US/docs/Web/HTTP/Status/402

You have depleted your monthly included credits. Purchase pre-paid credits to continue using Inference Providers. Alternatively, subscribe to PRO to get 20x more included usage.
Waiting 4 seconds...

Attempt 4/5 failed: HfHubHTTPError: Client error '402 Payment Required' for url 'https://router.huggingface.co/v1/chat/completions' (Request ID: Root=1-6aa53095-3c21006674815c3b37fc95cb;eb160a88-f740-4214-a4c2-d5d4594a2db9)
For more information check: https://developer.mozilla.org/en-US/docs/Web/HTTP/Status/402

You have depleted your monthly included credits. Purchase pre-paid credits to continue using Inference Providers. Alternatively, subscribe to PRO to get 20x more included usage.
Waiting 8 seconds...

Attempt 5/5 failed: HfHubHTTPError: Client error '402 Payment Required' for url 'https://router.huggingface.co/v1/chat/completions' (Request ID: Root=1-6aa5309d-08d31bf9566e7a3f1e0960d0;3acfe969-2ebe-493f-a792-b9a19b88b507)
For more information check: https://developer.mozilla.org/en-US/docs/Web/HTTP/Status/402

```

Figure 4.2.14: API Pilot execution

API failures that occurred in the pilot experiments are recorded in Figure 4.2.14. The system returns repeated HTTP 402 Payment Required errors from the Hugging Face inference route, and shows that the monthly credits are used up. All five attempts shown failed with payment-related errors, so those pilot requests were not completed (Achanta, 2025). These payment-

related failures reveal a limitation of API-based Generative AI experiments: inference requests may be interrupted by service availability or credit restrictions even after the connection has been tested successfully. This affects reproducibility, so methodological limitations must be distinguished from infrastructure failures. It also shows that failures should be recorded, not ignored.

	prompt_type	ground_truth	prediction	response_time_seconds	explanation
0	Zero-Shot	Unverified	Misleading/False	0.543119	No credible evidence or reputable reports conf...
1	Definition-Based	Unverified	Unverified	0.540176	No reliable sources confirm that U.S. agencies...
2	Few-Shot	Unverified	Unverified	0.628261	There is no verifiable evidence from reliable ...
3	Zero-Shot	Unverified	INVALID_LABEL	0.577127	
4	Definition-Based	Unverified	Unverified	0.797485	
5	Few-Shot	Unverified	Unverified	0.617388	No verifiable records or widely reported news ...
6	Zero-Shot	Misleading/False	Unverified	0.714128	There is insufficient publicly available evide...
7	Definition-Based	Misleading/False	Reliable/Supported	0.622844	Multiple reputable news outlets reported that ...
8	Few-Shot	Misleading/False	Unverified	0.609936	There is no clear, verifiable evidence confirm...
9	Zero-Shot	Reliable/Supported	API_ERROR	NaN	
10	Definition-Based	Reliable/Supported	API_ERROR	NaN	
11	Few-Shot	Reliable/Supported	Unverified	0.658054	
12	Zero-Shot	Unverified	Unverified	1.320832	No reliable sources confirm the existence of s...
13	Definition-Based	Unverified	Unverified	0.499594	No credible evidence or reliable reporting con...
14	Few-Shot	Unverified	Unverified	0.591976	There is no publicly available, verifiable evi...

Figure 4.2.15: Pilot Experiment Results

The outputs from the pilot experiments are shown in Figure 4.2.15 for the various types of prompts. The results include the Zero-Shot, Definition-Based and Few-Shot prediction, with responses generally returned in less than one second. The outputs include Unverified, Misleading/False and Reliable/Supported classifications, as well as some unsuccessful API responses. These pilot outputs show that each prompting condition produced classifications through the full experimental pipeline. The unsuccessful API responses point to unreliable service operation, while the Unverified predictions show the model using the abstention category. These observations reinforce the point that testing and evaluation of Generative AI should be conducted through multiple, controlled tests and not simply by fluent responses.

Although the pilot is far too small to estimate performance, its outputs are informative. Of the 15 requests (five claims, each under the three prompting conditions), two failed with API errors and one returned an invalid label, leaving 12 valid classifications. Ten of these 12 were Unverified, and all seven correct predictions were cases whose ground truth was Unverified; none of the Misleading/False or Reliable/Supported claims was classified correctly. This points to a possible tendency to default to Unverified, which the full experiment should test, because

a model that abstains on almost every claim would appear cautious while offering little help with triage. The pilot also produced a concrete example of the explanation risk discussed in Chapter 2: under Definition-Based prompting, a claim labelled Misleading/False was classified as Reliable/Supported, with the explanation that multiple reputable news outlets had reported it. Three responses returned no explanation at all. These observations are illustrative only, but they show why the balance of predictions, the quality of explanations and invalid outputs must be tracked in the full evaluation.

4.3 Key Findings

The implementation produced several significant insights into the application of Generative AI to emergency rumour classification. First of all, there were no empty source-text fields and a set of 1,972 records was successfully extracted through the PHEMEPlus extraction process. Each record included event information, thread IDs, tweet IDs, source text, claims, PHEME labels, research labels and annotation information. The extracted data therefore provided a sufficient basis for the experimental evaluation. The implementation also confirmed that all five emergency-event directories were recognised, so the final dataset covers multiple emergency scenarios rather than a single event.

A second finding is related to the allocation of the three research labels. There were 1,008 Reliable/Supported records, 393 Misleading/False records and 571 Unverified records. About 51.1% of the records were Reliable/Supported, 29.0% Unverified and 19.9% Misleading/False, so the original distribution is clearly imbalanced. The experimental procedure therefore drew 100 records from each category, giving a balanced sample of 300 observations. This gives parity among the three classes and reduces the risk that the largest class is over-represented in the comparative assessment (Guo et al., 2023).

There is a third finding which concerns evidence availability. Of the 1,972 records, 1,848 had evidence URLs and 124 did not. About 93.7% of the records therefore had an evidence URL. This is significant because here the research differentiates between Reliable/Supported, Misleading/False and Unverified information based on evidence and established ground truth, not on emergency relevance.

Table 4.3.1: Key Dataset Findings

Measure	Result	Analytical significance
Total rumour records	1,972	Provides the main empirical dataset
Reliable/Supported	1,008 (51.1%)	Largest research-label category
Misleading/False	393 (19.9%)	Smallest category
Unverified	571 (29.0%)	Important category where evidence is insufficient
Balanced experimental sample	300	100 records per category
Evidence URL available	1,848 (93.7%)	Most records have evidence references
No evidence URL	124 (6.3%)	Indicates some limitations in evidence availability
Emergency events	5	Provides cross-event variation
Repeated evaluations	3	Enables consistency assessment

Fourthly, representation differed substantially between emergency events: there were 522 threads in Sydney, followed by 470 in Ottawa, 458 in Charlie Hebdo, 284 in Ferguson, and 238 in Germanwings. The model is therefore evaluated across different emergency events rather than a single scenario.

Lastly, the pilot confirmed that Zero-Shot, Definition-Based and Few-Shot prompting could produce model outputs, some in less than a second. However, the pilot also revealed an

important operational constraint: HTTP 402 Payment Required errors kept appearing from the Hugging Face inference route because the monthly credits had been exhausted. The available findings indicate that the pipeline was implemented and the pilot executed successfully, but they cannot be used to confirm final accuracy, Macro-F1, precision or recall values for the 300-record experiment. The pilot did, however, suggest a possible tendency to default to the Unverified label (10 of 12 valid pilot predictions), which the full experiment should examine.

4.4 Critical Analysis

The findings show that the experimental pipeline was implemented successfully, but also that Generative AI reliability cannot be assessed based on successful response generation only. The main strength is the well-prepared structure of the PHEMEPlus dataset, including the extraction of 1,972 observations without missing source-text values. In addition, each record has been assigned to one of the three research categories, which facilitates comparing the model outputs with the ground-truth classifications.

Balanced sampling plays a key role in the research. The initial dataset includes many more Reliable/Supported observations than Misleading/False ones. Had the original class distribution been used, overall accuracy might have been biased towards the majority class. Balancing the categories by taking 100 records from each of them ensures an equal number of observations within each of the three categories. However, this sampling design means that the experimental sample does not reflect the actual distribution of the full PHEMEPlus dataset.

Table 4.4.1: Critical Interpretation of Experimental Evidence

Dimension	Evidence from implementation	Critical interpretation
Dataset quality	1,972 records; no blank source text	Strong basis for systematic testing
Class balance	100 records per category	Improves class comparison but differs from original distribution

Evidence availability	1,848 URLs; 124 without URLs	Strong evidence coverage, but not universal
Prompt conditions	Zero-Shot, Definition-Based, Few-Shot	Allows comparison of instruction effects
Repeated testing	Three repeats	Supports consistency assessment
API operation	API test successful	Confirms initial model connectivity
Pilot execution	Outputs generated under all three prompts	Demonstrates practical feasibility
API failure	HTTP 402 during pilot	Indicates operational limitations
Full performance metrics	Not reported in available results	Prevents unsupported claims about final accuracy

The evidence availability is another important issue. While 1,848 records have evidence URLs, 124 do not. This allows Unverified claims to be treated separately, rather than forced to be either true or false. This aligns with the conceptual framework set out in Chapter 2, in which cases with insufficient evidence are treated as uncertain, handled with caution and, where needed, verified by humans.

The pilot results also show a difference between technical functionality and research reliability. There was a successful connection with the API and the pilot generated responses for all three prompting strategies. But HTTP 402 failures illustrate that other external infrastructure and resource availability can affect model-based experiments. This is important because incomplete API execution may leave fewer observations for comparison and bias conclusions drawn from the remaining responses (Karuppanasamy et al., 2026).

Another key factor is prompt variation. The method aims to include Zero-Shot, Definition-Based and Few-Shot prompts alongside semantically similar prompt variations. This is meant

to establish whether different wording affects classification judgements when the meaning is unchanged. This testing is crucial because the literature shows that fluent-model responses do not necessarily indicate reliable classifications (Altayeva, 2025). Classifications must therefore be judged in terms of consistency, prompt sensitivity, uncertainty and error characteristics.

Most importantly, complete numerical results for accuracy, precision, recall, Macro-F1, false-positive rate, false-negative rate or confusion matrices are not found in the available Chapter 4 evidence. These metrics have been defined in the methodology and at present the results provided reflect the preparation of the dataset and the pilot behaviour and not the full experimental output. There are therefore no credible grounds for asserting a specific accuracy for the model or for judging one prompting technique superior. The findings do, however, show that the experimental setup was adopted, the data were properly processed, and the pilot demonstrated both successful model operation and practical API limitations.

4.5 Summary

This chapter presented the implementation outcomes and a critical analysis of the Generative AI evaluation on the PHEMEPlus dataset. A total of 1,972 rumour records were extracted across five emergency events, including 1,008 classified as Reliable/Supported and 393 as Misleading/False, as well as 571 Unverified records. Because the original dataset was imbalanced, 100 observations per category were used in the experiment.

The quality assessment of the data revealed that there were no missing values of the source-text and 1,848 of the 1,972 records (93.7%) had an evidence URL, while 124 (6.3%) did not. The implementation connected successfully to the chosen model, openai/gpt-oss-120b, through the API and produced pilot outputs under Zero-Shot, Definition-Based and Few-Shot prompting. Three repeated evaluations were built in to explore consistency and prompt sensitivity.

The pilot phase also uncovered HTTP 402 Payment Required API failures, an operational limitation that affected the intended testing. Importantly, the available results do not yet include complete final classification metrics. The results thus demonstrate the quality and viability of the experimental pipeline and stress the importance of obtaining full model-output results before making quantitative statements about the capabilities of Generative AI.

Chapter 5: Conclusion and Recommendations

5.1 Conclusion

The dissertation examined the trustworthiness of Generative AI in identifying false and misleading public-safety information in emergency scenarios. It examined how well AI can distinguish Reliable/Supported, Misleading/False and Unverified information, and how its performance should be judged in terms of consistency, prompt sensitivity, hallucinations, unsupported explanations, and false positives and negatives.

The study was carried out on the PHEMEPlus dataset, which includes 1,972 rumours from five emergency events: Charlie Hebdo, Ferguson, Germanwings, the Ottawa shooting and the Sydney siege. These comprised 1,008 Reliable/Supported records, 393 Misleading/False records and 571 Unverified records. A balanced sample of 300 records, 100 from each category, was chosen for the experiment.

An important conclusion is that Unverified should not be merged with Misleading/False. During an emergency, information may be incomplete or not yet available, and absence of evidence is not evidence of falsehood. Keeping this category separate allows more cautious verification of information.

Three prompting methods, Zero-Shot, Definition-Based and Few-Shot, were set up with openai/gpt-oss-120b and designed to be repeated to analyse prompt sensitivity and consistency. The planned quantitative measures were accuracy, precision, recall, Macro-F1, false-positive and false-negative rates and confusion matrices, supported by qualitative analysis of hallucinations and safety-related misidentifications.

However, the numbers required to complete all performance measures were not available in Chapter 4. No accuracy percentages can therefore be reported, and no definitive advantage of any prompting strategy should be claimed.

The pilot showed that the API functioned under the planned testing process, although some requests failed because of service or payment restrictions. The small pilot also hinted at a tendency to default to the Unverified label, which future runs should test. Overall, the evidence suggests that Generative AI can help process information and support triage, but should not be

treated as a decision-maker. Where evidence is limited, or where an incorrect classification could have serious public-safety implications, human verification remains necessary.

5.2 Recommendations

1. Use Generative AI under human supervision

Generative AI ought not to establish the veracity of safety-critical information itself but should assist processes for information triage and identification of potentially problematic claims (Komendantova and Erokhin, 2025). Uncertain or high-risk classifications should be reviewed by human analysts.

2. Retain the Unverified category

Verification systems should be able to recognise when the available evidence is inadequate. Insufficient evidence should not be forced into a “true” or “false” answer.

3. Integrate external evidence

Future systems should combine AI-driven classification with independent evidence, context and verification. This is especially relevant because evidence was present for 1,848 records in the dataset and absent for 124.

4. Test prompt sensitivity

Prior to using Generative AI for analysing emergency information, organisations should test semantically similar prompts (Miller, 2024). If the classifications differ significantly, they should be reviewed further.

5. Maintain complete audit records

Model identifiers, prompts, input claims, outputs, classifications, explanations and experimental configurations should be preserved (Zeng, 2023). This documentation supports reproducibility and makes it easier to investigate wrong or conflicting results.

5.3 Limitations of the Study

There are a number of limitations in the study. First, although 1,972 records were available in the PHEMEPlus dataset, only 300 of those were included in the principal experiment. Second,

the research was limited to five emergency events and one Generative AI model, which limits generalisation to other model types, events and platforms. Third, evidence was not uniformly available for every record (Ibrahim et al., 2026). The generalisability of the findings should be judged with these factors in mind.

A further limitation is API availability. The pilot reported failed API calls, reflecting how infrastructure and the availability of services can impact the continuity of Generative AI experiments. Future studies should therefore differentiate between technical failure and true model performance failure and report the distinction transparently.

5.4 Practical Implications

The research is relevant to organisations involved in emergency communication, crisis monitoring and public-safety information analysis. Large volumes of social-media data could be processed by a Generative AI system to detect claims that warrant further investigation (Vicari et al., 2026). However, the results must be presented as an analytical aid, not as undisputed facts. Any information that is Unverified, inconsistent or potentially harmful should be escalated for human review.

The study also recommends putting in place clear workflows and processes to record outputs from the model, inspect the supporting information and escalate high-risk data. These would help to reduce over-reliance on AI-generated explanations and ensure accountability if Generative AI is part of emergency-information processes.

5.5 Future Research

Further experiments could use a larger number of observations drawn from more emergency datasets. Inclusion of multiple models would allow for the comparison of model-specific behaviour and generate more evidence of generalisability (Huey et al., 2026). Different languages, social-media platforms and contemporary emergency situations should also be studied to test whether the findings hold.

Future research should include the calculation and comparison of the full set of performance measures as outlined in this dissertation (namely the accuracy, precision, recall, Macro-F1, false-positive and false-negative rate). Repeated experiments should also establish whether classifications remain consistent across similar prompts and repeated model calls.

Future research could also build an automated human-in-the-loop escalation system that refers Unverified, inconsistent and high-risk outputs to a human analyst. This would take the current research a step further, towards a safer operational system for verifying emergency information.

5.6 Final Summary

In conclusion, this dissertation has proposed a framework for assessing the reliability of Generative AI in emergency information analysis. The research relies on evidence-based labels, balanced sampling, controlled prompts, repeated assessment and error analysis, rather than on model fluency or accuracy alone.

Its contribution is an evaluation framework that takes into account the interdependence of accuracy, uncertainty, consistency, prompt sensitivity and human verification. Generative AI could assist in the process of emergency information analysis; however, it is important to have proper evidence, transparency of the process and human involvement where necessary.

References

- Achanta, A., 2025, July. Synthetic Lies: Security and Ethical Challenges of Deceptive Content in Generative AI. In 2025 3rd World Conference on Communication & Computing (WCONF) (pp. 1-6). IEEE.
- Al-Shamrani, S.H. and Hasan, R., 2025, September. Enhancing Public Safety in Smart Cities Through LLM-Based Misinformation Detection. In *2025 IEEE World Forum on Public Safety Technology (WF-PST)* (pp. 238-243). IEEE.
- Altayeva, A., 2025. A survey on intelligent sound event detection systems for public safety and emergency response. In: Proceedings of the 11th International Scientific Conference «Academics and Science Reviews Materials», Helsinki, Finland, 9–10 October 2025. *188p* (p. 42). University of Applied Sciences.
- Anggraini, D., Ulandari, D. and Sekti, B.A., 2026. Challenges of Social Sensing From a Safety and Security Perspective. In *Advanced Approaches for Trust and Identity Management in AIoT Environments* (pp. 211-244). IGI Global Scientific Publishing.
- Badu, J., Kruke, B.I. and Saetren, G.B., 2023. Crisis communication and trustworthiness among crisis actors: towards a typology of crisis management difficulties. *Safety in Extreme Environments*, 5(2), pp.119-130.
- Bajwa, A., 2025. AI-based emergency response systems: A systematic literature review on smart infrastructure safety. *Available at SSRN 5171521*.
- Buoncompagni, G., 2026. “Virality Alert”: The Construction, Imagination, and Algorithmic Falsification of a Local Disaster. *Journalism and Media*, 7(1), p.58.
- Chen, X., 2025. Sustainable Agile Identification and Adaptive Risk Control of Major Disaster Online Rumors Based on LLMs and EKGs. *Sustainability*, 17(19), p.8920.
- Chowdhury, A., 2026. Agentic AI for National Resilience: A Framework for Autonomous Information Verification in Critical Infrastructure. *Available at SSRN 6103406*.
- El Mikati, I.K., Hoteit, R., Harb, T., El Zein, O., Piggott, T., Melki, J., Mustafa, R.A. and Akl, E.A., 2023. Defining misinformation and related terms in health-related literature: scoping review. *Journal of Medical Internet Research*, 25, p.e45731.

Elhosary, M., Issaka, B., Mohammed, I., Sissy, B. and Raheem, S.A., 2026. Communicating the Crisis or Misinforming the Public? Analyzing the AI-Generated Visual Rhetoric of the 2025 California Wildfires. *International Journal of Communication*, 20, pp.1147-1166.

Guo, D., Chen, H., Wu, R. and Wang, Y., 2023. AIGC challenges and opportunities related to public safety: A case study of ChatGPT. *Journal of Safety Science and Resilience*, 4(4), pp.329-339.

Halford, E., 2026. Police application of generative artificial intelligence: A scoping review. *Journal of Policing, Intelligence and Counter Terrorism*, pp.1-30.

Hocevar, K.P., 2024. Who drives disaster communication? an analysis of twitter network structure and influence during a wildfire crisis. *Western Journal of Communication*, 88(4), pp.840-860.

Hu, S., Zhou, J., Tang, L., Zhao, Y. and Xing, H., 2024. The impact of risk communication, trust, and coping appraisal on individual preparedness decisions in geological hazard-prone areas. *International Journal of Disaster Risk Reduction*, 100, p.104166.

Huey, L., Ferguson, L., Heffren, C. and Furlano, M., 2026. “We are Not Ready for That”: Police Views on the Use of Generative AI to Create Public Security Threats. *CrimRxiv*.

Ibrahim, A., Ojebuyi, B., Fola-Adebayo, T., Makinde, A., Egbokhare, O., Aborisade, P., Olowoniyi, J., Eegunlusi, T., Adebomi, O., Okere, O. and Oguntade, F., 2026. Expert Perspectives on Detecting Fake News and Misinformation Governance Using Generative Artificial Intelligence in Nigeria: A Qualitative Exploratory Study. *Feedback International Journal of Communication*, 3(1), pp.74-87.

Jambholkar, M., 2024. Ethical Foresight: Confronting Misinformation, Representation and Toxicity in Generative AI.

Karuppanasamy, K., Appavu, K., Han, J., Lee, S.C., Kim, Y.W. and Moon, S., 2026. AI-Based Social Media Analytics for Disaster Resilience: A Unified Framework for Reliable Information Extraction. *KSCE Journal of Civil Engineering*, p.100643.

Kashyap Pargi, M., Abraham, A. and Chandra Nagavarapu, S., 2025. Transforming Public Safety and QoL in Smart Cities: Automated Event Detection and Response Generation with AI

and IoT. In *Internet of Vehicles and Computer Vision Solutions for Smart City Transformations* (pp. 267-290). Cham: Springer Nature Switzerland.

Komendantova, N. and Erokhin, D., 2025. Artificial intelligence tools in misinformation management during natural disasters. *Public Organization Review*, 25(1), pp.81-105.

Kumar, S., Sai, S., Chamola, V., Gaur, A., Agarwal, C., Huang, K. and Hussain, A., 2025. Peeping into the future: Understanding and combating generative AI-based fake news. *Cognitive Computation*, 17(3), p.103.

Li, Y. and Sheng, D., 2024. Determinants of public emergency information dissemination on social networks: A meta-analysis. *Computers in Human Behavior*, 152, p.108055.

Lu, J., Zhang, H., Xiao, Y. and Wang, Y., 2024. An environmental uncertainty perception framework for misinformation detection and spread prediction in the COVID-19 pandemic: artificial intelligence approach. *JMIR AI*, 3, p.e47240.

Miller, S.K., 2024. *Emerging Technology in Emergency Management: Using generative artificial intelligence (AI) to improve emergency messaging* (Doctoral dissertation, Capitol Technology University).

Mukherjee, C., 2026. AI-Based Detection of Deepfakes and Misinformation on Social Media. *Euro Vantage journals of Artificial intelligence*, 3(2), pp.9-30.

Park, S. and Nan, X., 2026. Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *AI & SOCIETY*, 41(2), pp.1501-1515.

Phillips, J. and Chan, E., 2025, May. Monitoring & Responding to Information Ecosystem Incidents: A Conceptual Framework Understanding Information Ecosystem Risk and Leveraging Academic Expertise for Information Incident Response. In *Proceedings of the International ISCRAM Conference* (Vol. 22).

Shen, H., Ju, Y. and Zhu, Z., 2023. Extracting useful emergency information from social media: A method integrating machine learning and rule-based classification. *International Journal of Environmental Research and Public Health*, 20(3), p.1862.

Shoaib, M.R., Wang, Z., Ahvanooey, M.T. and Zhao, J., 2023, November. Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. In *2023 international conference on computer and applications (ICCA)* (pp. 1-7). IEEE.

Skumanich, A., Kim, H.K., Sawyer, N., Davariar, R., Hill, J.D. and Kan, M., 2025, October. AI Monitoring of Social Media to Assist in Crisis Preparedness and Response: COVID-19 Case Study. In *2025 IEEE Global Humanitarian Technology Conference (GHTC)* (pp. 140-147). IEEE.

Stewart, I.S., Sevilla, E., Barragán, K. and Menteşe, E.Y., 2023. Disaster risk communication requires dissemination, dialogue and participation. *Nature Reviews Earth & Environment*, 4(12), pp.805-806.

Stolero, N., Elkady, S., Labaka, L., Verlin, J., Branlat, M. and Adini, B., 2024. Using social media in disaster management: The perceptions of emergency responders versus the public. *Risk, Hazards & Crisis in Public Policy*, 15(2), pp.128-161.

Tomassi, A., Falegnami, A. and Romano, E., 2024. Mapping automatic social media information disorder. The role of bots and AI in spreading misleading information in society. *PLoS ONE*, 19(5), p.e0303183.

Truong, D., 2026. AI-Enabled Generative Learning for Early Warning of Safety-Critical Small UAS Incidents under Sparse Data Conditions. *Journal of Safety Science and Resilience*, p.100362.

Tsang, S.J., 2026. Misinformation, disinformation, and fake news? Proposing a typology framework of false information. *Journalism*, 27(3), pp.719-739.

Vicari, R., Komendantova, N., Elroy, O., Dallo, I., Rapaport, C., de Carolis, C. and Yosipof, A., 2026. A toolbox to deal with misinformation in disaster risk management. *AI & SOCIETY*, 41(4), pp.3819-3843.

Wang, Y.Z., Xiao, Y., Yue, L., Gao, W., Du, Y., Fang, P., Di, S. and Zhang, M.L., 2026. AI-generated Images Challenge Visual Trust in High-risk Scenarios. *arXiv preprint arXiv:2607.22745*.

Wardle, C. and AbdAllah, A., 2023. The information environment and its influence on misinformation effects. *Managing infodemics in the 21st century: addressing new public health challenges in the information ecosystem*, pp.41-51.

Yang, S., Chung, H., Singh, D. and Shams, S., 2023. A two-step approach to detect and understand disinformation events occurring in social media: A case study with critical times. *Journal of Contingencies and Crisis Management*, 31(4), pp.826-842.

Zeng, L., 2023. Generative AI in public opinion guidance during emergency public events: Challenges, opportunities, and ethical considerations. SSRN Working Paper, 30 March 2023.

Zhai, W., Yu, H. and Song, C.Y., 2024. Disaster misinformation and its corrections on social media: spatiotemporal proximity, social network, and sentiment contagion. *Annals of the American Association of Geographers*, 114(2), pp.408-435.

Zhang, J., Kodama, M., Wu, Z., Chen, M., Zhu, Y. and Hong, G., 2026, June. Frontier Safety Policies for AI Emergency Preparedness in China. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency* (pp. 956-973).

Zhang, Y., Xie, Y., Shi, V. and Yin, K., 2023. Dynamic characteristics and evolution analysis of information dissemination theme of social networks under emergencies. *Behavioral Sciences*, 13(4), p.282.

Image citation

Aïmeur, E., Amri, S. and Brassard, G., 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1), p.30.

Jadi, A., 2016. Improving the emergency services for accident care in Saudi Arabia. *International Journal of Advanced Computer Science and Applications*, 7(9).

NVIDIA, n.d. What is generative AI and how does it work? [online] Available at: <https://www.nvidia.com/en-us/glossary/generative-ai/> [Accessed 12 September 2026].