The Role of Prior AI Experience in Evaluating Human vs. LLM-Generated Rationales for Disaster Tweet Classification

By

Cheng Chen

Bachelor's Thesis

Cheng.chen@ru.nl

S1102030

Radboud University

International Business Communication

Supervised by Jane Arleth dela Cruz

June 2025

Table of Contents

Abstract

The rapid growth of large language models (LLMs) has led to a phenomenon where various industries are using LLMs in workplaces. Rationales as reasonings behind AI-generated decisions play a critical role in building trust and transparency in human-AI collaboration. The current study aims to examine how does individual's prior experience with AI influence their perceived reliability and consistency of rationales in a disaster-related context. 114 participants completed the survey online via Qualtrics, where they were presented with the tweets, classifications, and rationales. No significant interaction effect among prior AI experience, type of rationales, perceived reliability and consistency was found, suggesting people with different levels of prior AI experience do not evaluate the reliability and consistency of the rationales differently, regardless of the generators. An unexpected significance was that people with high prior AI experience evaluated the consistency of the rationales more positively than those with low prior AI experience. These findings suggest that, in disaster contexts, LLM-generated rationales are perceived as equally reliable and consistent as human ones, supporting the viability of integrating LLMs into urgent tweet classification systems to enhance efficiency and real-time response. Future research should explore contextual factors that may shape human perceptions of LLM-generated content.

Key words

Large Language Model, Artificial Intelligence, classifications, rationales, disaster tweets, reliability, consistency.

Introduction

Social media serves not just as a leisure activity but also as a critical platform for saving lives in natural disasters. With the help of e-WOM (electronic word of mouth), people can send emergency messages faster than ever. A paper by Dugdale et al. (2012) showed the importance of Twitter (now X) in aiding organizations in rescue during the 2010 Haiti earthquake. Over 1.5 million tweets flooded the platform with requests for help and updates. The researchers found that Twitter, alongside SMS, became a lifesaving tool for people to request aid, providing rescuers with real-time information about trapped individuals.

However, an enormous amount of data could challenge the information filtering process, leading to difficulties in identifying urgent and legitimate emergency tweets (Plotnick & Hiltz, 2016). The Haiti earthquake did not only show how social media could be a lifesaving tool during disasters but also revealed the practical hardship in processing massive information in urgent times. The director of a crisis-mapping platform indicated they were not capable of handling such an immense amount of data (Forum, 2013). This calls for a rapid and accurate information classification system to process online information during disasters. While the classification of disaster-related content can effectively filter actionable posts for crisis response (Kruspe et al., 2021), human classification systems still face inherent limitations. Professional emergency responders often struggle with constraints such as the lack of trained staff to effectively monitor social media in real time (Seneviratne et al., 2024).

With the growth of Large Language Models (LLMs) in recent years, many studies have suggested using LLMs to classify disaster tweets (Lei et al., 2025). LLM is a type of machine learning model that uses a massive number of parameters to understand and generate human languages. LLMs can be trained to understand the tweets and classify them into different categories. A key element of the classification process is rational, which refers to the annotations used as the reference standard for explanations (Resck et al., 2024). In the context of tweet classifications, rationales are

the extracted parts of the tweet to justify the reason behind a certain classification. Adopting LLMs to The remaining question is how consistent and reliable the rationale of LLMs is compared to human-generated rationales. The current study is based on the evaluations by participants who were exposed to both LLM-generated rationales of disaster tweet classifications and human-generated rationales of disaster tweet classifications. However, considering the nature of the participants, they are very likely to be familiar with the use of LLMs (ChatGPT), which might produce bias when evaluating LLMs. The current study will also take prior experience with the use of LLMs into account.

Literature review

Importance of classifications

Disaster tweet classification involves categorizing tweets related to current disasters to provide faster response and allocate resources. Singh et al. (2019) built a classification system to identify emergency tweets during disasters. This system showed overall satisfying results, as it successfully identified more than 80% of the emergency tweets and correctly classified 76% of the cases. This system also effectively labeled emergency tweets with high priority. This result demonstrated the effectiveness of classification and its potential in filtering cases with high or low priority and more possible expansions.

A study by Stowe et al. (2016) applied a similar approach to explore the effectiveness of a tweet classification system. This system could filter and categorize tweets into relevant and fine-grained classes such as preparation, evacuation, sentiment, and reporting. Experts provided the annotations for the system. The findings revealed a high precision in identifying and classifying tweets compared to a simpler classification system. This paper further emphasized the importance of a complex and effective classification system, as it can yield better performance.

The current paper adopted the classification method from Alam et al. (2021). This paper introduced a large-scale (contains 77,000+ labeled tweets from 19

disasters), human-labeled dataset of disaster-related tweets for humanitarian aid classification. The humanitarian aid categories in this dataset form the basis for the current paper's tweet classification labels. The current paper categorized the tweets into five classifications:

1) Injured or dead people

2) Infrastructure and utility damage

3) Rescue, volunteering, or donation effort

4) Caution and advice

5) Sympathy and support

The selection was based on the presence of these categories across various disaster events.

Rationales in classification

The classification system helps to label the disaster-related tweets more effectively, and understanding the reasonings behind such classifications helps to add layers of transparency (Nguyen & Rudra, 2022). The reasoning behind classifications, known as rationales, has been shown to enhance classification performance when integrated into the classification system. This is supported by the core of explainable AI (XAI), which refers to the methods that aim to make AI decisions more understandable and transparent. This approach is essential in dire circumstances, such as law or health-related applications (Gohel et al., 2021). Ghaffarian et al. (2023) selected 68 publications to explore the applications of XAI in critical situations, where they detected a significant increase in the use of XAI in disaster management.

Sharma and Bilgic (2018) conducted a study on the effect of rationales in classifications on documents. They incorporated the rationales into the training process of document classification models and compared the results with models with no rationales. The findings showed that classification models with rationales significantly outperformed the models without rationales. Models with rationales had

higher levels of AUC (area under the curve) than models without rationales, indicating that the rationales helped the classifier differentiate between classes more effectively.

A hierarchical approach was conducted by Zhang et al. (2016), who compared the performance of a model with rationales (RA-CNN) against several powerful models without rationales. The researchers applied the models across five datasets to classify text and movie reviews. It was shown that RA-CNN outperformed all other models without rationales in every dataset. RA-CNN performed with higher accuracy and was more effective when it came to long documents. These findings emphasize the significant improvement rationales bring to the original models.

Human rationales

To dive deeper into the effect of rationales in machine learning, we start from analyzing human-generated rationales. Based on previous studies that suggested human rationales can improve the accuracy of text classification tasks, Strout et al. (2019) further proved the findings. The researchers found that models with human rationales outperformed the models without human rationales in terms of both accuracy and quality. They also discovered that the former was preferred by human judges over the latter. The findings are in line with Arous et al. (2021), who pointed out that existing models without human rationales may wrongly identify unrelated elements from the input when classifying texts. They proposed a system that incorporated human rationales, and it was shown to outperform the system without human rationales in terms of accuracy and interpretability. Together, these studies stressed the significant role of human rationales in enhancing machine learning models.

Inconsistent Performance of LLM-Generated Rationales

It was concluded that having rationales can enhance the quality of the classifications, but human-generated rationales in classification tasks often face time-consuming difficulty and are not scalable for real-time crisis events (Nguyen &

Rudra, 2024). With the development of LLMs, studies have proposed to incorporate LLM-generated rationales into classification tasks (Mishra et al., 2024; Pangakis & Wolken, 2024). These studies tended to compare the performance between human-generated rationales and LLM-generated rationales to justify this proposal. A study by Leinonen et al. (2023) investigated the effectiveness of code explanations generated by students versus LLMs like GPT-3. The research involved approximately 1000 first-year programming students who were asked to create and evaluate code explanations for three functions. The findings revealed that LLM-generated rationales were rated significantly higher in terms of accuracy and understandability compared to student-generated rationales. The study suggests that LLM-generated rationales outperformed human-generated rationales, indicating the strength of LLMs and the possibility of integrating LLMs into practice.

The paper by Wiegreffe et al. (2022) investigated the capability of LLM-generated explanations in classification tasks, particularly their consistency, factual accuracy, and alignment with human rationales. The study found that GPT-3 explanations are often preferred over crowdsourced ones due to their grammaticality and fluency, which is in line with a study by Mishra et al. (2024), who also found that the LLM rationales were preferred over existing crowdsourced rationales.

However, LLM-generated rationales also face significant limitations, as suggested by mixed findings from multiple studies (Rimban, 2023; Xu et al., 2024). A study by Kunz and Kuhlmann (2024) explored the limitations and challenges of LLM-generated rationales and overall performance in detail. They found that LLMs such as ChatGPT often imitate human-generated rationales and try to be more humanlike. This tendency has both advantages and drawbacks. Being humanlike shows potential great task-solving skills, but also the imperfections of humans. The two most significant traits were selectivity and illustrative elements, with subjectivity and misleading being less frequent. This behaviour could be potentially problematic because LLMs may choose humanlike traits over faithfulness and trustworthiness.

The study by Nguyen and Rudra (2024) investigated how LLMs perform in generating rationales for important information in crisis-related tweets compared to human annotations. The study found that ChatGPT's annotations were less precise than human annotations, with slight differences. The authors propose an active learning-based interpretable classification model to reduce the burden of manual annotations by selecting and annotating the most informative tweets. Experiments on disaster datasets show that ChatGPT can effectively extract rationales. Wiegreffe et al. (2022) discovered that LLMs lack depth and specificity, making them less reliable in complex reasoning tasks. These findings collectively emphasize the necessity for further research on the LLM-generated rationales.

When it comes to clinical reasoning tasks, LLMs face severe inflexibility issues. Kim et al. (2025) found that LLMs often yielded rationales that showed a lack of common sense, overconfidence in incorrect answers, and high hallucination rates. It was suggested that LLMs had limited generalization ability and lacked a logical capacity when it came to unfamiliar, out-of-distribution situations. This is alarming for disaster-related tweet classification, considering the unexpected nature of disasters.

The mixed findings underpin the strengths and limitations of LLM-generated rationales. LLM-generated rationales can increase the classification and overall task-solving speed, but also face systematic flaws. This further provide the necessity of the current study.

Evaluations

It is important to evaluate the characteristics of LLM-generated rationales and compare them to human-generated rationales because this provides a clearer view of the actual performance of LLMs. This paper chose to examine the reliability and consistency of both human and LLM-generated rationales.

Reliability

In this paper, reliability refers to trustworthiness (Martell et al., 2025), quality, accuracy, and practical applicability. These metrics were reflected in the questionnaire in multiple items. A high reliability of the rationales of disaster-tweet classification indicates that the rationales successfully reflect the reasoning of the classification, and people can feel reliably and trustworthily adopt the classifications in real-life situations. A low reliability refers to inaccurate reasoning, and the participants would not feel confident to allow this system to classify disaster tweets in practice.

Even human-generated rationales are often treated as the standards (Herrewijnen et al., 2024), they still face criticism. A study found that human-generated rationales can be spurious, as they spread bias and discrimination, resulting in low reliability (Sen et al., 2021). When compared to LLM-generated rationales, humans did not perform better in terms of accuracy and sufficiency (Carton et al., 2020).

Aljohani et al. (2025) concluded that LLM-generated rationales face significant challenges in truthfulness, safety, privacy, etc., which could reduce the overall reliability of LLM-generated rationales. Another study by Liu et al. (2024) explored the trustworthiness by administering a comprehensive survey. The results suggest that LLM-generated content can gain higher trustworthiness if they have higher alignment with publicly claimed information. This finding aligns well with Wong and Paritosh (2022), who conducted a study on improving the reliability of LLM-generated rationales. The researchers found that by requiring consensus among multiple LLMs, the reliability of LLM-generated content was significantly enhanced.

Consistency

The current paper defines consistency as stability and predictability (Cerqueira et al., 2025), in which we aim to test how stable the rationales of the tweet classifications are and the possibility of trusting the rationales to predict future classifications.

Similar to the reliability, human-generated rationales can be seen as inconsistent (Carton et al., 2020). The researchers found that human-generated rationales face domain inconsistency issues across different datasets and models.

A study by Saxena et al. (2024) evaluates the consistency and reasoning capabilities of both public and proprietary LLMs. The researchers found that proprietary models generally outperform public models in terms of consistency and reasoning, though none achieved a score of 90% in both areas. The experiments revealed that inconsistencies in responses often accompany incorrect answers and rationales.

The study by Bonagiri et al. (2024) evaluated the moral consistency of LLMs across semantically equivalent scenarios. The researchers found that LLMs often perform moral inconsistencies, which can undermine their reliability and trustworthiness. The authors developed the Moral Consistency Corpus, comprising 50,000 moral questions and responses from 11 LLMs to assess consistency. The experiment revealed that current LLMs models have significant moral inconsistencies and a lack of reliability.

Overall, both human and LLM-generated rationales face challenges in reliability and consistency, indicated by the mixed findings. Human-generated rationales are often seen as the gold standard or the benchmark (Herrewijnen et al., 2024), but commonly result in varied quality and inconsistency. Researchers pointed out the flaws of LLM-generated rationales, but are also actively exploring possible improvements.

Prior experience with the use of LLMs

AI has been one of the most trending topics in the past decades and has increased even more dramatically in the past few years with the birth of ChatGPT. 86 % of students in higher education have used AI for their schoolwork (Veera Korhonen, 2025). This alarms us to take the influence of using AI into account in user studies.

This paper examines the influence of prior experience (both familiarity and attitudes) with the use of LLMs on the perceptions of LLM-generated rationales. A study by Petzel and Sowerby (2025) tested the effect of prior experience with AI (such as ChatGPT). They found that previous experience with LLMs could lead to greater trustworthiness and intentions of AI. This proved that familiarity with LLMs could affect one's perceptions. Similar findings were found by Bach et al. (2024), who found that positive experiences with AI lead to firm trust with the AI provider, meaning the trust can be expanded to other AI systems by this provider. This study also discovered that negative experiences can be overruled by the need to use AI. This further demonstrated that people with previous experiences and dependency on LLMs could have a bias toward LLMs. To conclude, previous (positive) experience with AI could produce significant preferences over AI, and the current paper will further explore the effect of the preference.

Research gap

Previous studies (Aljohani et al. (2025); Bonagiri et al. (2024); Carton et al. (2020) focused on comparing the performance between LLM-generated rationales and Human-generated rationales regarding accuracy, understandability, effectiveness, etc. in different contexts. However, little to no studies focused on the reliability of classification rationales in the context of disaster tweets while taking the influence of prior experience with LLMs into account. This study aims to fill this gap, focusing on the reliability and consistency of LLM-generated rationales and human-generated rationales, while considering the effect of potential bias from using LLMs.

Research Question: **How does the prior experience with the use of LLMs influence the judgment of LLM-generated rationales of disaster tweet classifications in terms of consistency and reliability compared to human-generated rationales?**

Hypothesis: Participants with high previous experiences with the use of AI score higher on reliability and consistency of LLM-generated rationales compared to

human-generated rationales, while participants with low prior AI experience score higher for human-generated materials.

The current study contributes to existing discussions on the performance of LLM-rationales at both the scientific level and societal level. From the scientific perspective, the current study fills the gap of a clear definition of reliability of LLM-generated rationales and further investigates the potential bias of previous experience with the use of LLM. From the societal perspective, this study helps us to better understand how the experience of AI, both the familiarity and attitudes, could impact our evaluations of them. Nonetheless, this study can improve the tweet classification system during disaster situations, aiming to help more people in need.

Method

Materials

The current study consisted of two independent variables. The first independent variable was the type of rationales, generated by humans and LLMs. There were 1200 tweets from X about a Philippine flood event in 2024 collected by the thesis supervisor. The supervisor also provided an annotation protocol, which was finalized by the thesis group (for the annotation protocol, see Appendix A). The LLM-generated rationales and classification were provided by the supervisor (Llama 3.2), and the human-generated rationales and classification were annotated by the thesis group.

The thesis group was formed by four bachelor's students, meaning each student annotated 300 tweets and second-coded a sample of a peer's data (30 tweets per student) to calculate inter-annotator agreement. The inner coding reliability among the thesis group members was poor: $\kappa = .39$, $p < .001$. The reliability of the difference between the annotation of LLMs and humans was also low: $\kappa = .30$, $p < .001$. This aligned well with the researcher's observation, where the LLMs often select the entire relevant sentence, while the humans tended to select the key words. There were also notable overlaps.

To select the final tweets for the survey, the LLMs and humans must agree on the classifications, but not the rationales. The rationales should correspond to the text in the tweets to be valid. A total of 50 valid sample tweets were randomly chosen for the survey. The survey contained five tweets that were randomly assigned from the sample tweets, along with the classifications and rationales. An example of the survey can be found in Appendix B.

The second independent variable was the previous experience with the use of LLMs (or AI, to be more specific). Participants reported their experiences with AI with seven-point Likert scales anchored by 'completely disagree – completely agree' following two statements 'I have frequently used AI (such as ChatGPT) to solve questions in the past two years'; 'My experience with AI has been satisfying' (based on Dawar et al. (2022); Petzel & Sowerby, (2025)). A post-hoc awareness question *'The classification and rationales you have seen are provided by either a human or Large Language Model (Artificial Intelligence software such as ChatGPT). Which one do you think you are being exposed to?'* was asked at the very end of the survey. This can further test if there are correlations between assumptions on the provider of the rationales and the overall evaluations.

Subjects

A total of 252 people joined the survey, and 114 sufficient responses were recorded. 57.9% of the participants were female. Out of the 17 nationalities, Dutch was the most frequent nationality (N = 54), followed by Chinese (N = 28) and Estonian (N = 6). Most of the participants had a bachelor's degree (N = 63), followed by a high school diploma or equivalent (N = 26) and a master's degree (N = 18). The participants' ages range from 18 to 55, with a mean of 26.59 years. The prior experience with AI was measured with a seven-point Likert scale. Participants were categorized into three groups based on their self-reported AI experience using a 7-point scale: those rating 1-4 were classified as having low experience, and 5-7 as high experience. The sample distribution revealed a strong skew toward high AI experience, with 89 participants (78.1%) falling in the high experience category, compared to only 25 (21.9%) with

low experience.

Age (t (110.84) = .25, p = .81), gender (χ^2 (3) = 4.10, p = .251), education (χ^2 (3) = 1.81, p = .614), and prior experience with AI (χ^2 (2) = 2.04, p = .36) were all equally distributed across human and LLM-generated rationales.

Design

The current study had a 2 (rationale: LLM-generated rationales and human-generated rationales) * 2 (prior experience with LLMs: low and high) between-subject experimental design.

Instruments

Participants filled out an online questionnaire where they evaluated the reliability and consistency of the rationales. Each questionnaire contained five tweets, the corresponding classifications, and rationales.

- Reliability will be measured with a seven-point Likert scale following the four statements 'The classification and rationale are reliable.'; 'I trust the rationales given to justify the classification.'; 'The rationales demonstrate a high level of credibility'; 'I am confident that the rationale in this tweet is accurate'. The reliability of the four items was high (α = .813).
- Consistency will be measured with a seven-point Likert scale following the three statements 'The rationales are stable across all the tweet classifications'; 'The rationales do not contradict each other when classifying tweets about the same type of disaster event'; The rationales follow a clear and predictable pattern when applied to different disaster-related classifications'. The reliability of the three items was acceptable (α = .691).

Procedure

The experiment was conducted as an online survey via Qualtrics, and every participant finished the survey individually. The survey was activated on 25 April 2025, and closed on 12 May 2025, lasting 18 days. The survey was completed by the

four IBC students, each having their own research focus. Only the answers relevant to the current study were be used. There was be a general information page regarding the purpose of the study and instructions, followed by a consent form. The questionnaire started with collecting the background characteristics, including gender, age, education, and experiences with the use of LLMs. Then, participants were presented with five disaster-related tweets, classifications of the tweets, and rationales for the classifications. After this, the participants answered questions related to the reliability and consistency of the rationales. There were two versions of the materials, namely by LLMs and humans, but the participants were not told which version they were being presented with before finishing the experiment. At the end of the survey, participants were told that the materials were made by either humans or LLMs, and they had to guess which one they were exposed to. After finishing the survey, there was a thank-you page indicating the end of the experiment. The survey took around ten minutes.

Statistical Treatment

The primary statistical tests for this study were chi-square, independent sample t-test, and two-way ANOVA. Chi-square was conducted to explore the distribution of the participants in different conditions, t-test was used to examine the evaluations of the reliability and consistency, and two-way ANOVA was conducted to test the interaction effect of prior experience with LLMs and the (potential) different evaluations between LLMs-generated rationales and human-generated rationales.

Results

The purpose of the current paper is to investigate the influence of prior experience of using LLMs on perceived reliability and consistency of classification rationales that humans and LLMs provide. This paper was conducted in a disaster-related context. The results of the main findings are presented in Table 1, and the correction rate of identifying the rationale provider is shown in Table 2.

Reliability

An independent sample t-test was conducted to test the relationship between perceived reliability and the type of rationale providers. No significant difference was found in the perceived reliability of the classification rationales between humans and LLMs (t (110.73) = .57, p = .57, 95% CI [-.27, .49]). Participants evaluated both human rationales (M = 4.6, SD = 1.1) and LLM rationales (M = 4.7, SD = 1.0) with medium ratings.

Consistency

Another independent sample t-test was conducted to examine the relationship between perceived consistency and the type of rationale providers. No significant main effect was found either (t (110.82) = -.12, p = .91, 95% CI [-.42, .37]). Similar to the evaluation of reliability, the ratings were average across the two conditions (Human: M = 4.8, SD = 1.1; LLM: M = 4.7, SD = 1.0).

Influence of Prior Experience with LLM

To find out if the participants' prior experiences in using AI can influence their perceptions of the rationale, a two-way ANOVA was performed. No significant interaction effect was found between prior experience and types of rationale provider on reliability (t (F (1, 110) = .43, p = .511) or consistency (t (F (1, 110) = .69, p = .410). A significant effect was found between consistency and prior AI experience (t (F (1, 110) = 8.01, p = .005, η^2 = .07). Participants with a high level of prior AI experience (M = 5.0, SD = 1.0) rated consistency more positive than participants with low prior AI experience (M = 4.4, SD = 1.1). Participants with high prior AI experience (M = 4.7, SD = 1.0) did not evaluate reliability differently from those with low prior AI experience (M = 4.5, SD = 1.0; t (F (1, 110) = 1.16, p = .284).

Table 1

Evaluations of Reliability and Consistency with Prior Experience in AI

	Human			LLM			Total		
	M	SD	n	M	SD	n	M	SD	n

Reliability									
Low AI Experience	4.4	1.2	20	4.7	1.0	19	4.5	1.0	39
High AI Experience	4.7	1.0	37	4.7	1.1	38	4.7	1.0	75
Total	4.6	1.1	57	4.7	1.0	57	4.7	1.0	114
Consistency									
Low AI Experience	4.3	1.3	20	4.5	1.0	19	4.4	1.3	39
High AI Experience	5.0	1.0	37	4.9	1.1	38	5.0	1.0	75
Total	4.8	1.1	57	4.7	1.0	57	4.8	1.1	114

Identification of classification and rationale provider

At the end of the survey, participants were told that the classification and rationales were provided either by humans or LLMs, and they needed to guess which one they were exposed to. The findings showed that 61.4% of the participants did not correctly identify if the materials were provided by LLMs or humans. Participants did not show significant differences between the two conditions when guessing the provider, as indicated by the chi-square test (χ^2 (1) = 2.37, p = .124).

Table 2

The Accuracy of Identifying the Material Provider

	Count	Percentage
Correct	44	38.6%
Incorrect	70	61.4%
Total	114	100%

To conclude, most participants (61.4%) struggled with differentiating between human and LLM-generated classifications and rationales. Participants who were exposed to human or LLM-generated content did not evaluate the reliability and consistency differently. Participants with high prior AI experience rated consistency significantly higher than those with low prior AI experience, regardless of whether rationales came from humans or LLMs. Reliability was evaluated equally across different levels of prior AI experience.

Discussion

This paper explored whether prior experience with AI affects how individuals evaluate the reliability and consistency of the classification rationales in a disaster-tweet context. The hypothesis that people with high prior AI experience would rate the rationales provided by LLMs more reliably and consistently, while people with low prior AI experience would evaluate the materials provided by humans more positively, was not confirmed.

Prior Experience in Influencing Reliability and Consistency

The main purpose of the current study was to examine how individuals' prior experiences with AI influence their evaluations of content provided by human/LLM. The results suggest that despite different levels of AI experiences, the participants did not assess the reliability and consistency differently, whether the rationales were generated by humans or LLMs. A clearer summary of the main findings is shown in **Figure 1**.

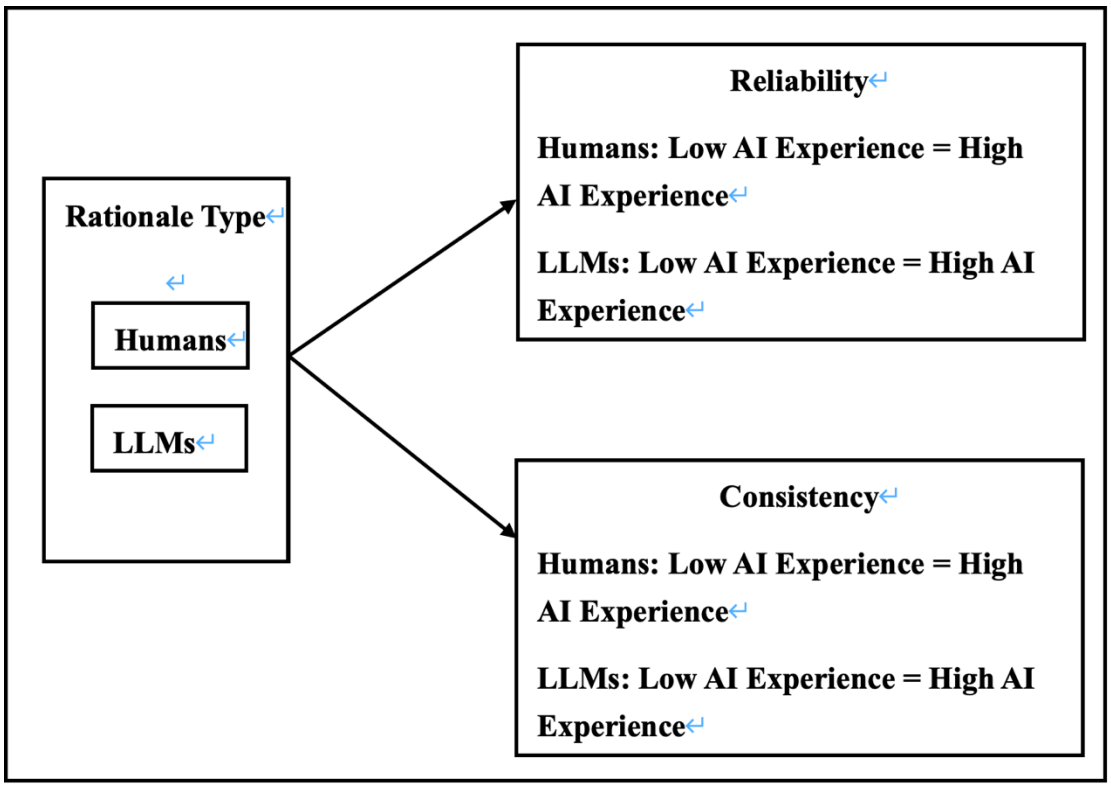

Figure 1. Summaries of the evaluation of reliability and consistency in function of the rationale type and prior AI experience ('=': no significant difference).

Starting with how participants evaluated the rationales. Overall, participants rated the reliability and consistency of the rationales with medium to high scores, indicating that both human and LLM-generated rationales for tweet classifications are rather reliable and consistent. This, however, contradicts previous studies (Aljohani et al., 2025; Bonagiri et al., 2024; Carton et al., 2020). These papers observed the unreliability or inconsistency of both human and LLM-generated rationales, in which Carton et al. (2020) also pinpointed the different performances between human and LLM-generated rationales. This contradiction is not entirely surprising, as it can be explained by the following marks.

Firstly, the comparison between the performance of human and LLM-generated rationales varies significantly in different conditions, as suggested by Chein et al. (2024). This paper compared the performance of human and LLM-generated rationales in three different tasks. The accuracy of the rationales varied drastically among the three tasks, yet humans and LLMs produced similar outputs. This finding

suggests that LLMs are human-like in the sense of imperfection, where humans' performance is based on context and content. In the current paper, the participants were exposed to a fixed content and setting. The evaluations may vary if the tweets are related to a different topic or the methodology is more complicated.

Secondly, the nature of the classification and its rationale may play a role in determining the outcome. Unlike generated videos or texts, classifications and the following rationales may reveal less originality, as the rationales were direct quotes of the tweets. This means that both rationales share similar or nearly identical lexical and structural features. As previously observed, when humans and LLMs agree on classification, they show overlap in the keywords that were selected. The agreement selection on classification could also potentially reduce the differences between human and LLM-generated rationales, and increase the reliability and consistency.

Lastly, most of the selected tweets are rather short, which can potentially reduce detectable variation between human and LLM rationale generation. This suggests that at least for the final selected tweets, humans and LLMs were not able to produce significant differences when annotating the rationales.

Moving to the influence of prior AI experience on perceived evaluations, which was found to be insignificant. Prior studies (Petzel and Sowerby, 2025; Bach et al., 2024) have found the opposite effect, where positive or high prior AI experiences lead to better evaluations of AI or even create bias. This conflict of significance can again be explained by the essence of tweet classifications and rationales. The rationales contain little information to reveal if they were generated by LLMs or humans, which was confirmed by the low correction rate when identifying the material provider. Participants' prior experience may have simply failed to contribute to the results. Another possible explanation could be the distribution of participants with different levels of prior AI experience. As mentioned, most of the participants self-reported a high level of experience with AI. This skewed distribution can potentially eliminate the potential significance that low prior AI experience can have. The findings suggest

that when evaluating rationales of online post classifications, participants' prior experience with AI does not influence the evaluations.

Additional findings

An unexpected significance was found between consistency and prior AI experience. Regardless of the type of rationale, participants with high prior AI experience evaluated the consistency of rationales more positively than participants with low prior AI experience. The cause of this effect is unclear, and no relevant studies have focused on the relationship between prior AI experience and the consistency of rationales in general. A possible assumption can be made to explore this phenomenon. People with higher prior AI experience may develop a mental model of AI reasoning. Oftentimes, chatbots such as ChatGPT provide responses based on the context and the user's preference, which can result in a more consistent and predictable pattern of responses (Funk et al., 2024). Participants who are frequently exposed to AI may subconsciously be more attuned to the flow and consistency of the rationales, leading to a more positive evaluation. This finding suggests that prior AI experience doesn't just affect evaluations for AI; it can also shape evaluative standards. As AI tools become more prevalent, it is important to understand how prior AI experience frames perceptions, even in situations where AI is not explicitly discussed.

The fact that the classification and rationale were provided either by humans or LLMs was revealed at the end of the study, and the participants were asked to identify which one they were exposed to. The main purpose was to investigate participants' abilities in differentiating human and LLM-generated rationales of classifications. A previous study mentioned that the correction rate of differentiating humans and LLMs varies across different media (Ramu et al., 2023). For example, people are better at identifying LLM-generated text than LLM-generated poems. Thus, it is no surprise that the majority of the participants failed to distinguish human and LLM-generated rationales that were extracted text from the tweets. The overall correction rate showed similarities to the current study.

Conclusion and Practical Implications

This study examined the influence of prior AI experience on individuals' evaluations of reliability and consistency of classification rationales in disaster-related content, as well as their ability to distinguish between human and LLM-generated rationales. Contrary to expectations, the hypothesis was not supported by the results. Participants with low and high prior AI experience evaluated the consistency and reliability of the rationales equally, with a medium to high score. Participants showed difficulty in accurately identifying the source of the classifications and rationales, regardless of their prior AI experience. The findings suggest that material complexity and participant characteristics may significantly influence the detectability of differences in human versus LLM-generated rationales.

The findings of the current study have a few implications. The type of media plays an important role in influencing perceptions. We can conclude that when the generated rationales are extracted from tweets or online posts in general, or when the rationale is short, people fail to distinguish between humans and LLMs, and the evaluations on the rationales do not differ. LLM-generated rationales show great alignment with humans when the generated content does not have the opportunity to reveal its originality. This highlights the ability of explainable AI in solving real-time issues. Emergency responders should consider integrating XAI into the online post classification system to enhance the task-solving speed.

It is also interesting to note the relationship between the consistency of the rationales and prior AI experience. This finding revealed that using AI can create unexpected mental preferences or focuses, even when AI is not being explicitly mentioned.

This study contributes to the growing literature on human-AI interaction, specifically in the domain of influence of prior experience with AI and evaluations of human and LLM output.

Limitations and Recommendations

A few limitations should be discussed to help future researchers conducting a more comprehensive study.

Despite the group effort in annotating and designing the survey, inevitably, the annotation process encountered unexpected difficulties and therefore took more time than scheduled. Each thesis member initially annotated 300 tweets and re-annotated them for the format requirement. The great workload could potentially degrade the annotation quality, which was confirmed by the poor internal reliability.

Furthermore, the majority of the participants were university students from the Netherlands, which fails to diversify the subject pool. Most of them share a similar background, especially the familiarity with AI. Having an insufficient number of participants with low prior AI experience can possibly dispel the significance it brings.

Lastly, the survey was more difficult to comprehend than expected. During the design process, the thesis group realized the potential comprehensive difficulties and repeated the meaning of rationales in the instructions. However, when collecting the survey, we received a certain amount of feedback stating that the survey was hard to follow. This is also reflected in the high unfinished rate of the survey. The participants who did finish the survey might not fully understand the questions, and therefore chose safe answers, as shown in the results.

Future research should take the difficulties of annotation, subject diversification, and survey hardship into account when designing the experiment. Besides reliability and consistency, future researchers can measure more characteristics of rationales to further investigate the differences between human and LLM-generated rationales. This can also be achieved by exploring different types of media, such as videos and long texts, or topics, such as crimes and entertainment.

Reference List

Alam, F., Qazi, U., Imran, M., & Ofli, F. (2021). HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 933–942. https://doi.org/10.1609/icwsm.v15i1.18116

Aljohani, M., Hou, J., Kommu, S., & Wang, X. (2025). *A Comprehensive Survey on the Trustworthiness of Large Language Models in Healthcare* (No. arXiv:2502.15871). arXiv. https://doi.org/10.48550/arXiv.2502.15871

Arous, I., Dolamic, L., Yang, J., Bhardwaj, A., Cuccu, G., & Cudré-Mauroux, P. (2021). MARTA: Leveraging Human Rationales for Explainable Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 5868–5876. https://doi.org/10.1609/aaai.v35i7.16734

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction*, *40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

Bonagiri, V. K., Vennam, S., Govil, P., Kumaraguru, P., & Gaur, M. (2024). SaGE: Evaluating Moral Consistency in Large Language Models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 14272–14284). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1243/

Carton, S., Rathore, A., & Tan, C. (2020a). Evaluating and Characterizing Human Rationales. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9294–9307. https://doi.org/10.18653/v1/2020.emnlp-main.747

Carton, S., Rathore, A., & Tan, C. (2020b). *Evaluating and Characterizing Human Rationales* (No. arXiv:2010.04736). arXiv. https://doi.org/10.48550/arXiv.2010.04736

Cerqueira, J. S. de, Kemell, K.-K., Rousi, R., Xi, N., Hamari, J., & Abrahamsson, P. (2025). *Mapping Trustworthiness in Large Language Models: A Bibliometric Analysis Bridging Theory to Practice* (No. arXiv:2503.04785). arXiv. https://doi.org/10.48550/arXiv.2503.04785

Chein, J., Martinez, S., & Barone, A. (2024). *Can human intelligence safeguard against artificial intelligence? Exploring individual differences in the discernment of human from AI texts*. In Review. https://doi.org/10.21203/rs.3.rs-4277893/v1

Dawar, S., Savita Panwar, S. P., Dhaka, S., & Kudal, P. (2022). Antecedents and Role of Trust in Chatbot Use Intentions: An Indian Perspective. *Marketing and Management of Innovations*, *13*(4), 198–206. https://doi.org/10.21272/mmi.2022.4-18

Dugdale, J., Van De Walle, B., & Koeppinghoff, C. (2012). Social media and SMS in the haiti earthquake. *Proceedings of the 21st International Conference on World Wide Web*, 713–714. https://doi.org/10.1145/2187980.2188189

Forum, S. W. (n.d.). *Crisis Maps: Harnessing the Power of Big Data to Deliver Humanitarian Assistance*. Forbes. Retrieved 1 March 2025, from https://www.forbes.com/sites/skollworldforum/2013/05/02/crisis-maps-harnessing-the-power-of-big-data-to-deliver-humanitarian-assistance/

Funk, P. F., Hoch, C. C., Knoedler, S., Knoedler, L., Cotofana, S., Sofo, G., Bashiri Dezfouli, A., Wollenberg, B., Guntinas-Lichius, O., & Alfertshofer, M. (2024). ChatGPT's Response Consistency: A Study on Repeated Queries of Medical Examination Questions. *European Journal of Investigation in Health, Psychology and Education*, *14*(3), 657–668. https://doi.org/10.3390/ejihpe14030043

Ghaffarian, S., Taghikhah, F. R., & Maier, H. R. (2023). Explainable artificial intelligence in disaster risk management: Achievements and prospective futures. *International Journal of Disaster Risk Reduction*, *98*, 104123. https://doi.org/10.1016/j.ijdrr.2023.104123

Gohel, P., Singh, P., & Mohanty, M. (2021). *Explainable AI: Current status and future directions* (No. arXiv:2107.07045). arXiv. https://doi.org/10.48550/arXiv.2107.07045

Herrewijnen, E., Nguyen, D., Bex, F., & Van Deemter, K. (2024). Human-annotated rationales and explainable text classification: A survey. *Frontiers in Artificial Intelligence*, *7*, 1260952. https://doi.org/10.3389/frai.2024.1260952

Kim, J., Podlasek, A., Shidara, K., Liu, F., Alaa, A., & Bernardo, D. (2025). *Limitations of Large Language Models in Clinical Problem-Solving Arising*

from Inflexible Reasoning (No. arXiv:2502.04381). arXiv. https://doi.org/10.48550/arXiv.2502.04381

Kunz, J., & Kuhlmann, M. (2024). *Properties and Challenges of LLM-Generated Explanations* (No. arXiv:2402.10532). arXiv. https://doi.org/10.48550/arXiv.2402.10532

Lei, Z., Dong, Y., Li, W., Ding, R., Wang, Q., & Li, J. (2025). *Harnessing Large Language Models for Disaster Management: A Survey* (No. arXiv:2501.06932). arXiv. https://doi.org/10.48550/arXiv.2501.06932

Leinonen, J., Denny, P., MacNeil, S., Sarsa, S., Bernstein, S., Kim, J., Tran, A., & Hellas, A. (2023). Comparing Code Explanations Created by Students and Large Language Models. *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 124–130. https://doi.org/10.1145/3587102.3588785

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2024). *Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment* (No. arXiv:2308.05374). arXiv. https://doi.org/10.48550/arXiv.2308.05374

Martell, M. J., Baweja, J. A., & Dreslin, B. D. (2025). Mitigative Strategies for Recovering From Large Language Model Trust Violations. *Journal of Cognitive Engineering and Decision Making*, *19*(1), 76–95. https://doi.org/10.1177/15553434241303577

Mishra, A., Rahman, S., Mitra, K., Kim, H., & Hruschka, E. (2024). Characterizing

Large Language Models as Rationalizers of Knowledge-intensive Tasks. *Findings of the Association for Computational Linguistics ACL 2024*, 8117–8139. https://doi.org/10.18653/v1/2024.findings-acl.484

Nguyen, T. H., & Rudra, K. (2022). Towards an Interpretable Approach to Classify and Summarize Crisis Events from Microblogs. *Proceedings of the ACM Web Conference 2022*, 3641–3650. https://doi.org/10.1145/3485447.3512259

Nguyen, T. H., & Rudra, K. (2024). Human vs ChatGPT: Effect of Data Annotation in Interpretable Crisis-Related Microblog Classification. *Proceedings of the ACM Web Conference 2024*, 4534–4543. https://doi.org/10.1145/3589334.3648141

Pangakis, N., & Wolken, S. (2024). Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels. *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, 113–131. https://doi.org/10.18653/v1/2024.nlpcss-1.9

Petzel, Z. W., & Sowerby, L. (2025). Prejudiced interactions with large language models (LLMs) reduce trustworthiness and behavioral intentions among members of stigmatized groups. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2025.108563

Plotnick, L., & Hiltz, S. R. (2016). Barriers to Use of Social Media by Emergency Managers. *Journal of Homeland Security and Emergency Management*, *13*(2), 247–277. https://doi.org/10.1515/jhsem-2015-0068

Ramu, D., Jain, R., & Jain, A. (2023). *Generation Z's Ability to Discriminate Between*

AI-generated and Human-Authored Text on Discord (No. arXiv:2401.04120). arXiv. https://doi.org/10.48550/arXiv.2401.04120

Resck, L. E., Raimundo, M. M., & Poco, J. (2024). Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales. *Findings of the Association for Computational Linguistics: NAACL 2024*, 4190–4216. https://doi.org/10.18653/v1/2024.findings-naacl.262

Rimban, E. L. (2023). Challenges and limitations of ChatGPT and other large language models. *International Journal of Arts and Humanities*, *4*(1), 147–152. https://doi.org/10.25082/IJAH.2023.01.003

Sen, I., Samory, M., Flock, F., & Wagner, C. (2021). *What Human Rationales Tell Us about Machine Explanations*.

Sharma, M., & Bilgic, M. (2018). Learning with rationales for document classification. *Machine Learning*, *107*(5), 797–824. https://doi.org/10.1007/s10994-017-5671-3

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2019). Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, *283*(1), 737–757. https://doi.org/10.1007/s10479-017-2522-3

Stowe, K., Paul, M. J., Palmer, M., Palen, L., & Anderson, K. (2016). Identifying and Categorizing Disaster-Related Tweets. *Proceedings of The Fourth International Workshop on Natural Language. Processing for Social Media*,

1–6. https://doi.org/10.18653/v1/W16-6201

Strout, J., Zhang, Y., & Mooney, R. (2019). Do Human Rationales Improve Machine Explanations? *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 56–62. https://doi.org/10.18653/v1/W19-4807

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., & Choi, Y. (2022). *Reframing Human-AI Collaboration for Generating Free-Text Explanations* (No. arXiv:2112.08674). arXiv. https://doi.org/10.48550/arXiv.2112.08674

Wong, K., & Paritosh, P. (2022). k-Rater Reliability: The Correct Unit of Reliability for Aggregated Human Annotations. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2*, 378–384. https://doi.org/10.18653/v1/2022.acl-short.42

Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., & Cambria, E. (2024). *Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond* (No. arXiv:2306.09841). arXiv. https://doi.org/10.48550/arXiv.2306.09841

Zhang, Y., Marshall, I., & Wallace, B. C. (2016). *Rationale-Augmented Convolutional Neural Networks for Text Classification* (No. arXiv:1605.04469). arXiv. https://doi.org/10.48550/arXiv.1605.04469

Appendix A

Annotation Protocol (Humanitarian Aid Information) (version March 2025)

Annotation Rules:

1. **Understand the Tweet**

   First and foremost, it is essential to read the entire tweet and fully understand it before labelling the word or word groups in the tweet.

   Only provide classification label and word-level rationale annotation for a tweet of a disaster event.

2. **Complete the Tweet Annotation**

   Provide **ONE** classification label and its corresponding word/ group of words rationales. There should only be one label and corresponding one word/group of word rationales.

   Corresponding words/group of words for rationales are to be written in a separate column.

   If two or more labels apply: choose the class label based on severity and need for immediate response as ranked below:

   1. Injured or dead people
   2. Infrastructure and utility damage
   3. Rescue, volunteering, or donation effort
   4. Caution and advice
   5. Sympathy and support

   In some cases, the same relevant word is used more than once in a sentence. All the relevant words should be annotated regardless of the duplicates.

   If you labelled even a single word in a tweet, go through the whole sentence and

check if there is any other information that should be also labelled. Do not leave any relevant word unlabelled in a sentence.

3. **Certainty**

   You should annotate both label and rationales only if, from the text, you are **sure/certain** that the word(s) relates to a specific event. Otherwise, in case of any doubt as a human reader, do not label the tweet, leaving it blank. In every case, a clear clue indicating the relation to an event and one of the classifications should be found in the text.

4. **Be Concise**

   The number of words included as a rationale should be **as short as possible (no extra irrelevant info added) without the loss of information**. The rationale should be short but informative on why the tweet is classified into its respective category.

   Descriptive adjectives *(Descriptive adjectives modify nouns by providing additional information about their appearance, size, shape, color, condition, or other qualities)* should not be part of the rationale word groups if they are not adding relevant information.

   Articles (a, an, the in English) should not be included into the rationale word groups.

5. **Label the quantitative adjectives**

   Quantitative adjectives are accepted as informative and should be included in the

   labelled word groups. These include *about, more than, less than, some, a few,*

   at least, at most and more.

6. **Punctuation**

   Include punctuation marks that change or add sentiment of the tweet to the rationale annotations (For example, it is acceptable to skip period, commas,

colon/semi-colon; but question/exclamation mark should be included)

Classification Label

- **Injured or dead people**: Reports of injured or dead people due to the disaster
- **Infrastructure and utility damage**: Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles
- **Rescue, volunteering, or donation effort**: Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, people in shelter facilities, donation of money, or services, etc.
- **Caution and advice**: Reports of warnings issued or lifted, guidance and tips related to the disaster
- **Sympathy and support**: prayers, thoughts, and emotional support

Rationales

Rationales are considered as words/phrases inside the tweet that are responsible for the tweet to be classified into a classification label.

- **Injured and dead people:** words or phrases that refer to the total number of injured, dead or killed during the disaster event. If an adverb indicates additional information of the number of people killed, this adverb should be labelled together with the number e.g.: " approximately", "about", "more-or-less", "more than", "less than", etc. Qualitative descriptions such as "several", "few", "some", etc. should also be included.

Example:

`18 dead bodies found in #Everest base camp #Earthquake #Nepal`

`RT @USER: LIVE: Strong aftershock hits India,a Nepal once`

again; death toll rises to 2,123 URL

- **Infrastructure and utility damage**: words or phrases of exact type of infrastructure and names of buildings, roads, utilities, transportation hubs, services should be included. The verbs/adjectives of the type and/or extent of damage it/they incurred should also be included such as 'destroy, devastated, collapse, ruin, close, open, etc.'

  Example:

  UNESCO World Heritage sites destroyed by quake: sad sight. All our thoughts and prayers are with the people of Nepal. URL

  RT @USER: Kathmandu's historic Dharahara tower collapses in Nepal quake, people reportedly trapped in rubble - pic via @USER

- **Rescue, volunteering, or donation effort**: words or phrases that describe the number of people affected and/or evacuated should be included and the exact verb/adjectives used should be included too such as "affected, missing, evacuated, stuck, stranded etc.'' Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, people in shelter facilities, donation of money, or services, etc.

  Example:

  Latest News: India Evacuates 170 Foreigners From Nepal URL

  RT @USER: Rain stops, rescue operations resume in #Kathmandu. #NepalEarthquake

- **Caution and advice:** words or phrases that report warnings issued or lifted, guidance and tips related to the disaster important from public safety and prevention of further casualties.

- **Sympathy and support**: words or phrases that contain prayers, thoughts, and emotional support.

  Example:

  RT @USER Some of the beautiful children I met in Nepal. Praying for their safety @USER URL

  @USER Thanks. Have those details. Our thoughts also with the people of Nepal.

Appendix B

Disaster Tweet Classification Survey

Information

We would like to invite you to participate in a research study. Participation is voluntary. Before you decide whether or not to take part, we will give you information about the study. Please take time to read the information carefully. If something is not clear, or you would like more information, please ask the researcher. The contact details can be found at the bottom of this letter.

Outline and aim of the research study

The study aims to investigate the classification procedure of tweets related to Philippine Flood Event that happened in 2024. The classification of tweets is based on the humanitarian aid information they contain, which includes 5 labels (Injured or dead people; Infrastructure and utility damage; Rescue, volunteering, or donation efforts; Caution and advice; Sympathy and support). Classification labels are supported by the extractive rationales, which can be seen as a set of words or phrases extracted from input texts that illustrate the decision-making process.

Who is conducting the research study?

This research is conducted by a BA Thesis group of International Business Communication students at Radboud University, the Netherlands. Sofiya Kuzmina, Cheng Chen, Martijn Keller, and Alicia Sao. The research is supervised by Jane dela Cruz and Eric Sanders.

What does participation in the study entail?

In this study, you will be given a block of 5 tweets that includes an initial tweet related to the Philippine Flood Event of 2024; a label of this tweet classification, and an extractive rationale that is highlighted in yellow color. First, you will answer a few demographic questions. Based on the block of tweets, you will be asked several questions. The survey is set to take approximately 10 minutes.

Voluntary participation

Your participation in this research is voluntary. This means that you can stop your participation and withdraw your consent at any time during the study, without giving a reason. Withdrawing from or stopping participation will not have any adverse consequences for you.

What data do we collect and how do we handle it?

In this study we collect research data. These will be used for scientific purposes. The research data will be anonymised. The research data may be of interest to other scientific research. The anonymised research data will therefore be available to other scientists for at least 10 years. Because the research data are anonymous , they cannot be traced back to you.

Do you have any questions?

If you have any questions or would like more information about the research study, please contact Sofiya Kuzmina (sofiya.kuzmina@ru.nl) Chen Cheng (cheng.chen@ru.nl) Martijn Keller (martijn.keller@ru.nl) Alicia Sao (alicia.sao@ru.nl) The research is supervised by Jane dela Cruz (janearleth.delacruz@ru.nl) and Eric Sanders (eric.sanders@ru.nl) For other questions about this study (e.g. data processing), please contact privacy@let.ru.nl

Consent

Clicking on the "I Consent" option below indicates that:

- You have read the above information
- You consent to participating in the research study as described in the above information
- You voluntarily agree to participate
- you understand that the research data will be available for at least 10 years for review and reuse in future scientific research.
- You agree that the data officer and designated data management staff of Radboud University may view your data. **-**
- you are at least 18 years of age

Consent

○ I have read and understand the information that is outlined to me above and I consent to participating in the study.

Q1 What is your gender?

- ○ Male
- ○ Female
- ○ Non-binary / third gender
- ○ Prefer not to say

Q2 How old are you?

Q3 What is your level of education?

- ○ High school diploma or equivalent
- ○ Bachelor’s degree
- ○ Master’s degree
- ○ Other

Q4 What is your nationality?

Instructions

(please read this through before answering the questions) On the next page, you will be presented with a block of five tweets regarding the Philippine Flood Event that happened in 2024. You will see some texts from the tweets are highlighted in yellow. These highlighted texts are **Extractive Rationales**, which explains the reasoning of the classification of the tweet. If a tweet contained two or more applicable labels, the final class label was selected based on the severity of the situation and the need for immediate response. The labels were prioritized according to the following ranking: 1. Injured or dead people 2. Infrastructure and utility damage 3. Rescue, volunteering, or donation effort 4. Caution and advice 5. Sympathy and support. Based on the block of tweets, we ask you a number of questions regarding the classifications and rationales. There are no right or wrong answers; therefore, be sure to answer questions honestly and based on your opinion. Extractive rationale - set of words or phrase in the input text extracted from input texts that illustrate the decision-making process of

why one of the five labels was assigned specifically to the presented tweet.

Tweet	Label
@Username 33 people were killed in landslides set off by Tropical Storm Trami in a province south of Manila, a Philippine police official said.	Injured or dead people
@Username The number of dead and missing in massive flooding and landslides wrought by Tropical Storm Trami in the Philippines has exceeded 100 and the president said Saturday that many areas remained isolated with people in need of rescue. **[URL Here]**	Injured or dead people
@Username Tropical Storm Trami triggers killer floods, landslides in Philippines and the death toll is mounting READ MORE: **[URL Here]** **[URL Here]**	Injured or dead people
@Username Oh please pray to everyone who are affected of typhoon Kristine here in the Philippines 馃槶馃槶 hope everyone who needs help already rescue and no more hurt. The wind is so strong and heavy rain is here again 馃槶 please stop this already 馃槶馃槶 it's extreme alert in my place. **[URL Here]**	Sympathy and Support
@Username Dear God, we lift up the Bicol region to you during tropical storm #Kristine. The heavy rains have caused severe flooding, forcing thousands of families to evacuate. We pray for their safety and comfort. With several provinces under Signal #2, we ask for your divine protection鈥 **[URL Here]**	Sympathy and Support

To what extent do you agree/disagree with the following statements based on the block of tweets above?

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
I trust the rationales given to justify the classification.	○	○	○	○	○	○	○
The classification and rationale are reliable.	○	○	○	○	○	○	○
The rationales demonstrate a high level of credibility.	○	○	○	○	○	○	○
I am confident that the rationale in this tweet is accurate.	○	○	○	○	○	○	○
The rationales do not contradict each other when classifying tweets about the same type of disaster event.	○	○	○	○	○	○	○

The rationales follow a clear and predictable pattern when applied to different disaster-related classifications.	○	○	○	○	○	○	○
The rationales are stable across all the tweet classifications	○	○	○	○	○	○	○