Evaluating Large Language Models for Confidence-based Check Set Selection

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown promise in automating high-labor data tasks, but the adoption of LLMs in high-stake scenarios continues to be a challenge due to two issues: their tendency to answer despite uncertainty and their difficulty handling long input contexts robustly. We investigate LLMs' ability to identify low-confidence outputs for human review through "check set selection"– a process where LLMs prioritize information needing human judgment. Using a case study on social media monitoring for disaster risk management, we define the "check set" as a list of tweets escalated to the disaster manager when the LLM has the least confidence, enabling human oversight within budgeted effort. We test two strategies for LLM check set selection: *individual confidence elicitation* – LLMs assesses confidence for each tweet classification individually, requiring more prompts with shorter contexts, and *direct set confidence elicitation* – LLM evaluates confidence for a list of tweet classifications at once, using less prompts but longer contexts. Our key contributions are: (1) we propose a novel performance metric for LLM-human collaboration in check set selection, (2) we compare individual and direct set-based selection strategies across input sizes and aggregation methods, and (3) we investigate LLMs' direct set selection capabilities from long-context inputs. Our results reveal that set selection via individual probabilities is more reliable but direct set confidence does show potential. Direct set selection challenges include such as inconsistent outputs, incorrect check set size, and low inter-annotator agreement. Despite these challenges, our approach improves collaborative disaster tweet classification, demonstrating the potential of human-LLM collaboration.

1 Introduction

Large language models (LLMs) have significantly advanced the field of natural language processing (NLP) and made it possible to automate a wide range of NLP tasks such as classification, information retrieval, summarization, and many more (Raiaan et al., 2024; Lee et al., 2022; Cohen et al., 2022; Yang et al., 2024). LLMs can perform these tasks by following prompts, where the enduser provides task details and input data, and the model generates a text response. However, studies show that endusers tend to struggle to identify incorrect LLM responses, a problem that can escalate as larger and more complex LLMs are less likely to refrain answering questions (Zhou et al., 2024).

The adoption of LLMs in high-stakes scenarios continues to be a challenge, as assuming LLM-generated responses to be always correct can have severe consequences, i.e., if incorrect outputs influence decision-making processes. Previous studies evaluated LLMs' ability to express uncertainty which we refer to as confidence elicitation (Xiong et al., 2024; Lin et al., 2022; Tian et al., 2023; Kadavath et al., 2022). Confidence elicitation methods have shown that uncertainty estimates are closely correlated with the accuracy of the prediction (Tian et al., 2023; Kumar et al., 2023). While LLM's output is impossible to evaluate automatically in the real-world setting, we investigate if we can surface LLM incorrectness using confidence elicitation techniques.

We introduce the check set for the human-LLM collaboration pipeline. The check set is list of potentially misclassified predictions by the LLM needing review by the endusers. It enables LLM and humans to work together by prioritizing areas where human judgment is most needed.

In this paper, we investigate the LLMs' check set selection capability with a case study in the field of disaster risk management. For this use case, the check set is a list of tweets escalated to the disaster manager when the LLM has the least confidence, enabling human oversight within a budgeted time-frame. LLMs have the potential to assist dis-

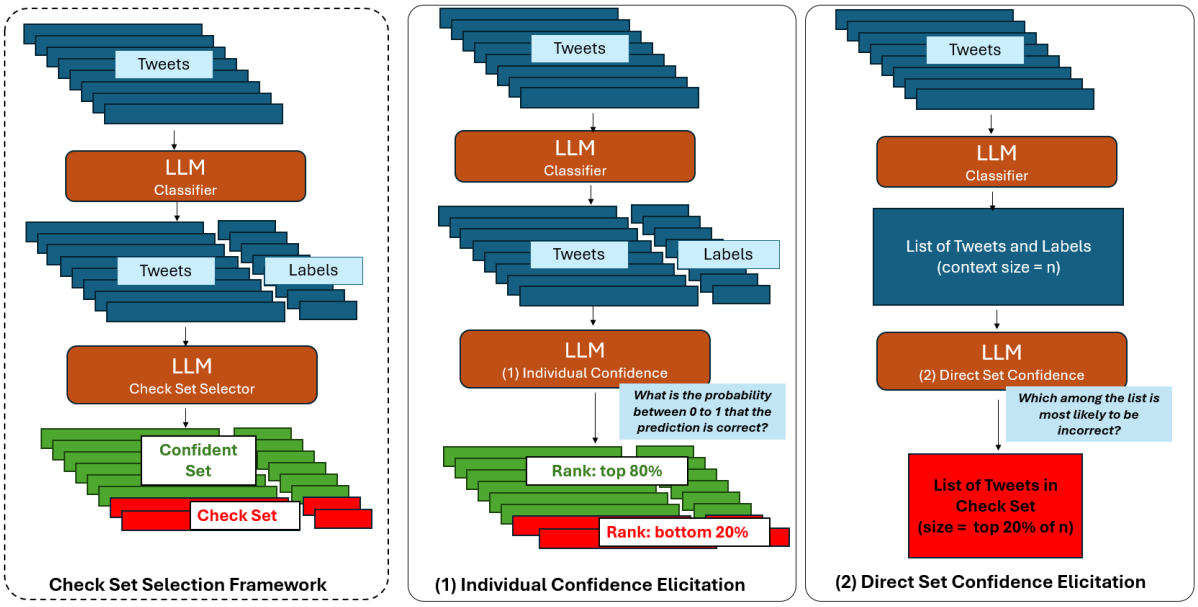


Figure 1: Check Set Selection Framework. Two strategies for check set selection (1) Individual Confidence Elicitation - LLM assesses confidence for each tweet classification individually, requiring more prompts with shorter contexts (2) Direct Set Confidence Elicitation - LLM evaluates confidence for a list of tweet classifications at once, using fewer prompts but longer contexts.

aster managers in sifting through massive amounts of online social media data for relevant, critical, and actionable information during disaster events. We present two methods for check set selection as seen in Figure 1: (1) *individual confidence elicitation*: LLM assesses confidence of each tweet classification separately using individual probabilities, requiring more prompts with shorter contexts and (2) *direct set confidence elicitation*: LLM evaluates confidence for a list of tweet classifications at once which allows for comparison within the list, using fewer prompts but longer contexts. These two approaches attempt to mitigate two underlying problems of LLMs in high-stakes use cases, LLMs refusing to refrain from answering questions they may not know the answers to (Zhou et al., 2024) and LLMs being unable to robustly make use of information in long input contexts (Liu et al., 2024).

Our key contributions are as follows:

1. We propose a novel performance metric for LLM-human collaboration in check set selection.

2. We compare individual and direct set-based selection strategies across input sizes and aggregation methods.

3. We investigate LLMs' direct set selection capabilities from long-context inputs.

While existing studies have investigated LLMs' ability to retrieve single information points or to make singular inference from a long-context (Hsieh et al., 2022; Gupta et al., 2024; Levy et al., 2024), investigating LLMs' ability to select a direct set of information points from long-context as input is under explored. Intuitively, more input data and long context provide LLMs more information i.e., the more classifications, the more comparisons LLMs can make to determine the potential incorrect classifications. However, recent studies show that LLMs struggle with long-context tasks, performing best when relevant information is at the start or end of the input and worse when it appears in the middle. (Liu et al., 2024; Hsieh et al., 2022).

We ran our experiments using both closed and open-sourced LLMs: gpt-4o-mini (OpenAI, 2024a), gpt-4o (OpenAI, 2024b), llama 3.1 8B (Llama Team, 2024), mistral 7B v0.3 (Jiang et al., 2023) across check set selection from predictions on two classification tasks: (1) humanitarian aid vs. not humanitarian aid and (2) humanitarian aid information type. Furthermore, we investigated the influence of different list-referencing methods and varying context-length.

Our results show that LLMs have the ability of check set selection using confidence elicitation techniques by outperforming random check set selection. Individual confidence elicitation is

found to be more reliable compared to direct set confidence selection. This is evidenced by issues in direct set method such as providing incorrect list sizes, inconsistent outputs across different list-referencing methods, and low inter-annotator agreement. However, we observe that direct set selection has potential and could be explored further as LLMs improve.

2 Method

The study focuses on the investigation of LLM's ability to select a useful check set from long-context input using confidence elicitation. First, we present the motivation of our approach and how we use LLMs as our disaster tweet classifiers. Then, we demonstrate the set selection methods. Lastly, we deep dive on the LLMs direct set selection ability from long context input.

Problem Definition. LLMs have been very effective in various natural language tasks. However, adoption of LLMs in high-stake scenarios continues to be a challenge due to two main issues: the larger and more complex the LLMs the less likely they are to refrain from answering questions they do not know the answer to and LLMs struggle with long-context tasks, having performance change with the position of relevant information. We aim to mitigate these problems using check set creation by allowing LLMs to utilize their confidence estimates of their initial predictions to prioritize information needing human review. We emphasize the need for LLM-human collaboration in these scenarios.

LLM as Disaster Tweet Classifier We test the performance of LLMs as disaster tweets classifiers using two classification tasks: Task (1) humanitarian aid vs. not humanitarian aid – asking LLMs if the tweet is useful for humanitarian aid or not and Task (2) humanitarian aid information classification – asking LLMs to classify the tweet based on the type of humanitarian aid information it contains. We ran our experiments on eight different disaster events, where each disaster event contains 100 tweets. More details are found in Section 3.1. The selected check sets are from the initially classified list by these classifiers.

Set Selection using Individual Confidence Elicitation. We make use of an LLM to predict the probability of the initial tweet classification from our disaster tweet classifier to be correct with a value between 0.0 and 1.0, referring to one of the methods by Tian et al. (2023) on confidence elicitation. We select the check set by using the tweet classifications with the lowest probabilities of being correct at the lowest 20% of the tweet classifications. The chosen check set size of 20% corresponds to the estimated effort the disaster managers have budget for, i.e., time and people to review check set. We chose a fixed check set size because it standardizes the effort done by the endusers and allows us to compare across different check set selection strategies. For cut-off tweets with the same probabilities, we use random selection.

Investigating LLMs Direct Set Selection Capabilities from Long Context. Given the list of tweets and classifications provided by an AI assistant, we prompt the LLM to identify the k tweets with potential erroneous classification labels. The task requires the LLM to understand the initial classification task prompt, access the list of k tweets and classifications, and use them to select the check set for the enduser. Figure 7 shows an example set selection prompt.

First, we investigate the influence of context length of the input so we ran prompts with different list context sizes of 25, 50, and 100 tweets and classifications. For the context size of 25 tweets and classifications, we divided the 100 tweets into 4 disjoint groups with each prompt selecting 5 from the list to create the check set size of 20. Second, we investigate the influence of referencing methods used for the tweet and classification lists. We do these investigations following Mizrahi et al. (2024)'s finding that instruction templates lead to very different performance. The four list referencing methods and their rationale are as follows:

- **numerical ID** – method commonly used for single retrieval from a list
- **full-text** – ensures LLM selects the actual tweets and not hallucinating IDs
- **keywords** – similar to how humans recall relevant information from a list of sentences
- **short-uuid** (8 characters) – used as key for single retrieval methods that is more robust than numerical IDs as hallucination can easily be detected.

We used multiple prompts (n = 10) for the same disaster event where in every prompt, we shuffled the order of the input list of tweet classifications randomly. This is to investigate whether or not the order influences the set selection choice. To select the final check set from the responses of the multiple prompts, we applied majority vote on valid responses.

3 Experimental Setup

3.1 Datasets

Task 1: humanitarian aid vs. not humanitarian aid. We randomly sampled 100 tweets for four different disaster events from CrisisBench (Alam et al., 2021b), a consolidated crisis-related social media dataset for humanitarian information processing. For the LLM prompt design, we renamed the class labels as *humanitarian aid* and *not humanitarian aid* from the original broad labels *informative* vs. *not informative* to explicate the labeling task.

Task 2: Humanitarian Aid Information Classification. For the humanitarian information classification task, we utilized human-annotated crisis-related tweets from (Alam et al., 2021a). The original dataset had 11 labels, however, we limited our labels to the 5 that were present in all of our selected crisis events, following (Zou et al., 2023) who also reduced their labels. Originally, we experimented with including the labels: *other relevant information* and *not humanitarian*, however, our initial experiments showed that such vague and negated labels are too challenging for the LLM. We sampled 100 tweets for each of the four different disaster events. More information about the datasets used is found in appendix A.2

3.2 Models

We chose four of the latest LLM's in our experiments. We used gpt-4o-mini (OpenAI, 2024a), gpt-4o (OpenAI, 2024b), llama 3.1-8B (Llama Team, 2024), and mistral 7B v0.3 (Jiang et al., 2023). These models were chosen because they are commonly used by both researchers and the public and have high capabilities in reasoning tasks. We ran our experiments at the temperature setting of 0.0 to make all models deterministic in their prediction. All the other parameters were kept default. The exact model parameters and information are found in Appendix A.3.1.

3.3 Prompts

Classifier Prompts. We formulated our classifier prompts with reference to the annotation protocol and the class description provided from the original dataset paper sources. We observed that choice of prompt strategies can influence the relative performance of the model which is in line with multiple works (Mizrahi et al., 2024; Wei et al., 2024; Gupta et al., 2024). So, we used the maximum performance metric of Mizrahi et al. (2024) to select the prompt templates used for our classifiers from different prompt strategies. The exact prompts can be found in the Appendix A

Individual Confidence Set Selection Template Prompts. The set selection prompts consists of the following: (1) individual confidence elicitation task, (2) the classification task prompt and (3) individual tweet and classification. We evaluated different prompt strategies for individual confidence elicitation from Xiong et al., 2024 and Tian et al., 2023 to find the best prompt strategy for our specific tasks. We used as our maximum performance metric (Mizrahi et al., 2024) effective accuracy to select our final prompt. Figure 6 shows the example individual confidence set selection prompt.

Direct Set Selection Template Prompts. The direct set selection prompts consists of the following: (1) the direct set selection task instruction, (2) the classification task prompt and (3) the list of k tweets and classifications. We manually craft the set selection prompt, where we make explicit the importance of the count of the items that need to be retrieved and that only items in the provided list are to be selected. The choice of prompt strategy also influenced the response here, so we again used maximum performance metric (Mizrahi et al., 2024). We used the most number of valid prompt response as our metric to select our final prompt. Figure 7 shows the example direct set selection prompt where the list-referencing method used was the full text.

3.4 Evaluation Metrics

First, we need to evaluate the initial performance of the LLM on classifying single tweets. We use the following metrics for this: **Accuracy** and **Effective Accuracy**. We define effective accuracy as the overall performance of the collaboration of the LLM and enduser on the dataset D of length n, when the enduser is provided with the set size of c to review. For this scenario, we are working with the assumption that the enduser's performance on the check set has 100% accuracy. This is computed as follows:

$$\%EffAcc_D = \%Acc_{LLM}\frac{(n-c)}{n} + \%Acc_{HUM}\frac{c}{n}$$

To evaluate the LLMs' ability to select a set from long context input, we introduce the following metrics:

No. of Valid Prompt Response. We test the

robustness of all the LLMs on their ability to provide valid prompt responses consistently. We count valid responses by the original long context input, i.e., by the 100 tweets input so 1 valid response is equivalent to 4 valid responses of each disjoint group of context size 25 and 2 valid responses of each disjoint group of context size 50. A response is considered valid if (1) the set provides the correct number of items requested and (2) all the items in the set come from the long-context input, i.e. there were no hallucinations.

Inter-Annotator Agreement. We used Krippendorff's alpha (Krippendorff, 1970) to measure the inter-annotator agreement between the multiple prompts with the varying classification list order.

4 Results

4.1 Disaster Tweet Classification Performance

We ran our experiments on two classification tasks across eight disaster events. The LLMs' performance for Tasks 1 and 2 are found in table 2 measured in accuracy scores at the column Acc. We observed that the closed-source models, gpt-4o-mini and gpt-4o perform well in both tasks, achieving accuracy scores of between 72% and 91% for Task 1 and between 84% and 95% for Task 2. Based on these accuracy scores, we observed that the chosen 20% check set size is the check size that would be needed for a good classifier, if the check set selection is perfect (see column *Eff Acc (Max)*, the maximum effective accuracies of the LLMs given the check set size in table 2. At the chosen check set size, the *Eff Acc (Max)* of almost all LLMs reach to above 0.85 across all tasks and all disaster events.

4.2 LLM Individual Confidence Check Set Selection Performance

Using the results from the initial classification tasks, we select our individual confidence check set based on the individual probabilities of each tweet classification of being correct. The effective accuracies of the different models for Tasks 1 and 2 are in table 2 using the individual confidence set selection strategy at column Eff Acc (I). All Eff Acc (I) is higher than the original accuracies of the models, hence improve overall classification performance.

To check the effectiveness of the individual confidence check set selection strategy, we compare Eff Acc (I) with the effective accuracy achieved by the models when selecting a random check set (column *Eff Acc (Random)*) of the same size. Note that there is a ceiling for effective accuracies as they are dependent on both the original accuracy and the chosen check size, we show these in column *Eff Acc (Max)*. We highlighted the instances where the individual confidence check set selection did not outperform random in table 2. We observed that for task 1, only gpt-4o's individual confidence check set selection outperformed random across all four disaster events, gpt-4o-mini's and mistral's outperform random most of the time, and only llama's fails to do so. However, for task 2, all the LLMs' individual confidence check set selection outperformed random across all four disaster events.

We wanted to know if there is an optimal check set size, compared to the current 20%, from our models by mapping the effective accuracies achieved by the models across changing check set sizes as seen in figure 8 in appendix A.5. These were the average effective accuracies from the four disaster events per task. We found that there is no obvious optimal check set size, with almost all models reaching 100% efffective accuracy only when all the tweets are checked.

4.3 LLM Direct Set Selection Performance

4.3.1 LLMs ability to select from a set is influenced by the input context size

As a first step to test LLMs' check set selection ability using direct set confidence elicitation, we count the number of valid prompt responses LLMs generate. Figure 2 shows the no. of valid prompt responses LLMs can generate by context size. We observed that the input context size influences some LLMs' ability to select a set from a list. We see this in figure 2 where llama is able to select from context sizes of 50 and 25 tweet classifications consistently over the larger context size of 100 using the short uuid referencing method for both tasks. Mistral, on the other hand, is able to consistently provide valid responses for context sizes 50 and 25 for only task 2.

4.3.2 The list-referencing method affects LLM's direct set selection output

Figure 3 shows the no. of valid prompt responses that LLMs can generate when asked to select 10 tweets from a list of 50 tweets and classifications by list-referencing method. We observed that the

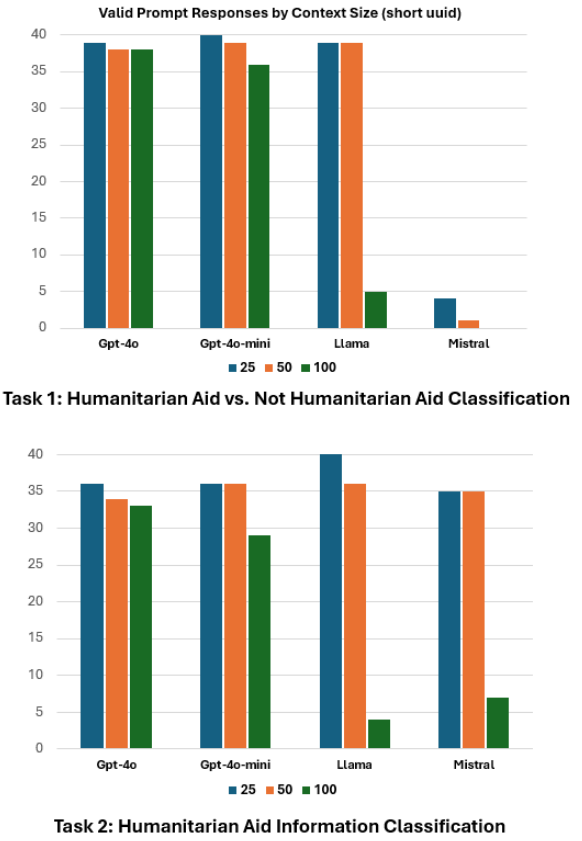


Figure 2: Valid Prompt Responses by Context Size using the short uuid referencing method.

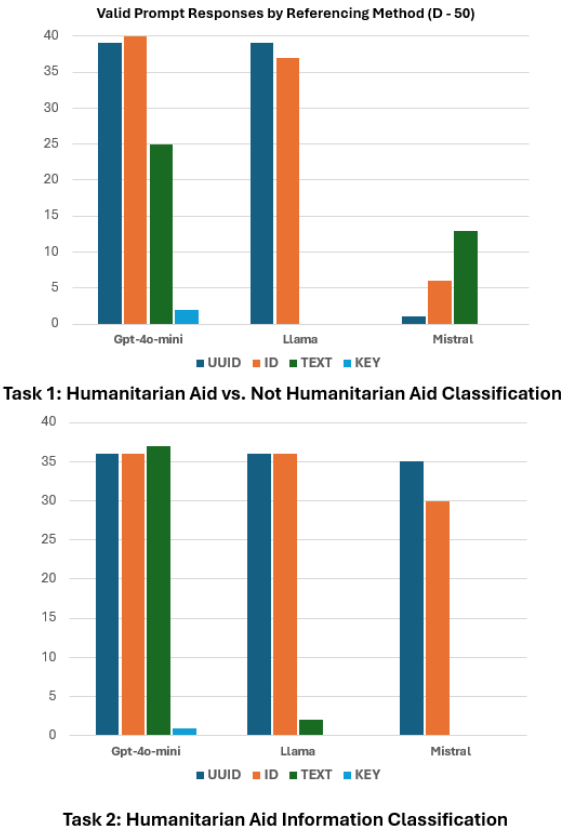


Figure 3: Valid Prompt Responses by Referencing Method at context size of 50 Tweets

chosen referencing method affects the no. of valid prompt responses generated. We observed that providing an index, i.e., either the ID or the short uuid in the list, helps LLMs retrieve a set from the input list. All LLMs struggled in retrieving the full tweet text and keywords, providing invalid responses as output, mostly providing incorrect number of tweets.

4.3.3 The input list order influences direct set selection.

We observed that the selected check sets vary significantly when we shuffle the order of the input list of tweets and classifications. We present the Krippendorf's alpha inter-annotator agreement scores for our models in Tasks 1 and 2 table 1 using the short uuid referencing methods. We do not have agreement scores for some models with insufficient valid prompts. The alpha is computed on the agreement across 100 tweets per disaster event i.e., whether they are included in the check set in each prompt iteration. We must take note that these agreement scores cannot be directly compared across context sizes but are to be evaluated individually. Table 1 shows that only gpt-4o and gpt-4o-mini had agreement scores above 0.50, with gpt-4o having 0.60 and above for two disaster events. This shows that input list order can influence the chosen check set using direct set selection.

4.4 Individual Confidence is more Reliable but Direct Set Confidence does show promise

The effective accuracies from the direct set confidence selection are shown in the columns Eff Acc (D - <context size>) in table 2. Effective accuracies for direct set selection across tasks and context sizes are higher than the original accuracies. We note that the effective accuracies for direct set sizes D-50 and D-25 are disadvantaged beforehand compared to the D-100, because they are dependent on the luck of the misclassified tweets being evenly distributed across subgroups. When compared with the effective acccuracies using random check set (*Eff Acc (Random)*), check set selection using gpt-4o outperforms random across all tasks and context sizes, gpt-4o-mini outperforms random at almost all events except Nepal Earthquake while llama and

Task 1: Humanitarian Aid vs. Not Humanitarian Aid				
Event	Model	D-100	D-50	D-25
California Earthquake	gpt-4o-mini	0.27	0.27	0.31
	gpt-4o	0.05	0.25	0.32
	llama 3.1	-0.06	0.03	0.22
	mistral v.03	-	-	-
India Floods	gpt-4o-mini	0.30	**0.59**	**0.56**
	gpt-4o	**0.55**	**0.55**	0.49
	llama 3.1	-	0.31	0.27
	mistral v.03	-	-	0.31
Nepal Earthquake	gpt-4o-mini	0.13	0.30	0.31
	gpt-4o	0.11	0.19	0.36
	llama 3.1	-	0.16	0.22
	mistral v.03	-	-	-
Vanuatu Cyclone	gpt-4o-mini	0.10	0.31	0.19
	gpt-4o	0.41	**0.55**	**0.63**
	llama 3.1	-	0.14	0.28
	mistral v.03	-	-	-
Task 2: Humanitarian Information Classification				
Event	Model	D-100	D-50	D-25
Mexico Earthquake	gpt-4o-mini	0.12	0.20	0.22
	gpt-4o	0.28	0.34	0.39
	llama 3.1	-	0.22	0.15
	mistral v.03	0.03	0.19	0.30
Sri Lanka Floods	gpt-4o-mini	0.36	0.38	0.44
	gpt-4o	0.40	**0.53**	**0.51**
	llama 3.1	-	0.30	0.37
	mistral v.03	0.13	0.34	0.40
Canada Wildfire	gpt-4o-mini	0.13	0.25	0.28
	gpt-4o	0.40	0.39	**0.60**
	llama 3.1	-	0.18	0.31
	mistral v.03	0.00	0.36	0.45
Hurricane Harvey	gpt-4o-mini	0.14	0.29	0.42
	gpt-4o	0.25	0.22	0.30
	llama 3.1	0.25	0.17	0.25
	mistral v.03	-	0.20	0.40

Table 1: Inter-annotator agreement between the valid prompts. Krippendorf's alpha by context size with short uuid referencing method

mistral have some events and context sizes that do not outperform random.

We compare the two check set selection strategies and observe that individual confidence check set selection is a more reliable method over direct set confidence selection for having insufficient valid responses. Effective accuracies from direct check set selection are higher than individual confidence check set selection for both gpt-4o and llama across all disaster events in task 1. Effective accuracies from individual confidence check set selection are higher than direct check set in all LLMs across almost all disaster events in task 2.

5 Discussion

We discovered from our experiments that although we set LLMs to their most deterministic setting, when we do direct check set selection, changing the order of the input context (list of tweets) lead to different check set selections and can even return invalid responses. This observation holds across different input context sizes. We recommend evaluating LLMs with multiple prompts always as we have observed that this is under reported.

6 Related Work

Confidence Elicitation in LLMs The most common ways to measure confidence in model predictions rely on model's internal logits. However, with the decoder-only LLMs, it has become less suitable to use these methods. There have been methods in prompting LLMs themselves to express uncertainty in natural language is referred to as verbalized confidence (Lin et al., 2022). Xiong et al. (2024) defines a systematic framework for LLM uncertainty estimation using prompting, sampling and aggregation strategies and benchmarks these methods in calibration and failure prediction. Tian et al. (2023) showed that large LLMs can express calibrated-confidence (as a probability or phrases like 'highly likely') more accurately than their raw conditional probabilities suggest.

LLM performance on long-context input text There are multiple studies that evaluate the long-context capabilities of LLMs (Hsieh et al., 2022; Shaham et al., 2023; Levy et al., 2024). Long-context" is an umbrella term for use cases of LLMs defined by the total length of the model's input that may include retrieval, summarization, and information aggregation (Goldman et al., 2024). The common task that papers evaluate on is the needle-in-a-haystack (NIAH) task, where the LLMs are tasked to retrieve the fact (the "needle") in a long input context (the "haystack") and asking the LLM to retrieve it given a related question (Kamradt, 2023). Hsieh et al. (2022) expands the NIAH task with a comprehensive evaluation of long-context LLMs by creating a new synthetic benchmark, RULER with flexible configurations length and task complexity. The paper revealed that almost all models exhibit large performance drops as context increases (Hsieh et al., 2022). Most papers evaluate LLM performance on synthetic datasets or existing benchmarks (Hsieh et al., 2022; Shaham et al., 2023; Levy et al., 2024), Gupta et al. (2024) differs by evaluating the gpt-4 suite of LLMs in solving progressively challenging tasks, as a function of factors such as context length, task difficulty, and position of needle using a created real-world financial news dataset.

Task 1: Humanitarian Aid vs. Not Humanitarian Aid								
Event	Model	Acc	*Eff Acc (Random)*	*Eff Acc (Max)*	Eff Acc (I)	Eff Acc (D-100)	Eff Acc (D-50)	Eff Acc (D-25)
California Earthquake	gpt-4o-mini	0.77	*0.82*	*0.97*	0.81	0.83	**0.85**	**0.85**
	gpt-4o	0.72	*0.78*	*0.92*	0.81	**0.84**	**0.84**	0.80
	llama	0.62	*0.70*	*0.82*	0.70	**0.72**	0.70	0.68
	mistral	0.67	*0.74*	*0.87*	0.75	-	**0.77**	-
India Floods	gpt-4o-mini	0.91	*0.93*	*1.0*	**0.96**	0.95	0.95	0.95
	gpt-4o	0.92	*0.94*	*1.0*	0.96	0.98	**0.99**	0.98
	llama	0.85	*0.88*	*1.0*	0.87	0.87	**0.93**	0.90
	mistral	0.83	*0.86*	*1.0*	0.86	-	-	**0.93**
Nepal Earthquake	gpt-4o-mini	0.78	*0.82*	*0.98*	**0.84**	0.81	0.81	0.80
	gpt-4o	0.77	*0.82*	*0.97*	0.83	0.84	**0.85**	0.83
	llama	0.77	*0.82*	*0.97*	0.81	**0.82**	0.81	0.80
	mistral	0.72	*0.78*	*0.92*	**0.79**	-	-	-
Vanuatu Cyclone	gpt-4o-mini	0.82	*0.86*	*1.0*	0.90	0.87	0.90	**0.91**
	gpt-4o	0.78	*0.82*	*0.98*	0.89	**0.92**	0.91	0.91
	llama	0.82	*0.86*	*1.0*	0.85	-	**0.86**	0.84
	mistral	0.73	*0.78*	*0.93*	0.77	-	-	**0.78**
Task 2: Humanitarian Aid Information Classification								
Event	Model	Acc	*Eff Acc (Random)*	*Eff Acc (Max)*	Eff Acc (I)	Eff Acc (D-100)	Eff Acc (D-50)	Eff Acc (D-25)
Mexico Earthquake	gpt-4o-mini	0.84	*0.87*	*1.0*	**0.92**	0.88	0.90	0.90
	gpt-4o	0.89	*0.91*	*1.0*	0.93	0.92	**0.95**	**0.95**
	llama	0.77	*0.82*	*0.97*	**0.84**	0.79	0.82	0.82
	mistral	0.68	*0.74*	*0.89*	**0.76**	0.71	0.74	0.73
Sri Lanka Floods	gpt-4o-mini	0.89	*0.91*	*1.0*	**0.92**	**0.92**	**0.92**	**0.92**
	gpt-4o	0.91	*0.93*	*1.0*	**0.95**	0.94	0.93	**0.95**
	llama	0.84	*0.87*	*1.0*	0.89	0.88	0.88	**0.90**
	mistral	0.73	*0.78*	*0.93*	**0.85**	0.78	0.80	0.78
Canada Wildfire	gpt-4o-mini	0.94	*0.95*	*1.0*	**1.0**	0.97	0.95	0.97
	gpt-4o	0.95	*0.96*	*1.0*	**1.0**	0.98	0.99	0.99
	llama	0.86	*0.89*	*1.0*	**0.95**	-	0.93	0.91
	mistral	0.84	*0.87*	*1.0*	**0.93**	0.87	0.85	0.85
Hurricane Harvey	gpt-4o-mini	0.86	*0.89*	*1.0*	0.90	0.89	0.90	**0.91**
	gpt-4o	0.85	*0.88*	*1.0*	0.90	0.89	0.90	**0.91**
	llama	0.71	*0.77*	*0.91*	**0.77**	**0.77**	0.75	**0.77**
	mistral	0.58	*0.66*	*0.78*	**0.70**	-	0.63	0.61

Table 2: Effective Accuracies of the Check Set Selection Strategies. *Eff Acc (Random)* is the effective accuracy for the task given a random check set, *Eff Acc (Max)* is the maximum possible effective accuracy for the task, Eff Acc (I) is for the individual confidence elicitation and Eff Acc (D) is for direct set confidence elicitation and the number indicates the context length size. The referencing method for direct set used for this table short-uuid

7 Conclusion

In this paper, we investigate the ability to identify low-confidence outputs for human review through check set creation, the process of utilizing LLMs to prioritize information needing human review. We run our experiments using a case study in disaster risk management. We tested two strategies for check set selection: individual confidence elicitation by assessing confidence for each tweet classification and direct set confidence elicitation by evaluating confidence for a list of tweet classifications at once. Furthermore, we examined LLMs' direct set selection capability by adjusting context sizes and list-referencing methods. Our results show that LLMs' struggle in direct set selection as they cannot consistently provide valid prompt responses such as incorrect list sizes and output information not found in the original input. Furthermore, we observed that direct set selection can be influenced by the list-referencing method, the input context size, and the list order of the input. Hence, we say that individual confidence set selection is more reliable than direct set selection for our particular setting. However, we observe that the direct set method has potential and could be explored further as LLMs continue to improve. Despite these challenges, our approach improves collaborative disaster tweet classification, demonstrating the potential of human-LLM collaboration.

8 Limitations

We only evaluated four commonly used LLMs: gpt-4o-mini, gpt-4o, llama and mistral. We only evaluated on the base models to test their check set selection capabilities. Instruction-tuning/fine-tuning these models to specifically do check set selection tasks may lead to more favorable results.

Our use case is focused on classification tasks for disaster risk management with text that are only in English language tweets. For the direct set confidence set selection, we only tested context sizes of 100, 50 and 25 tweets. A smaller context size may offer more stable responses from the LLMs. In addition, in selecting the check set from the smaller context sizes, D-50 and D-25, we did not try to optimize which tweets to compare with each other.

Our experiments were not performed in a real world application where we had an actual disaster manager perform the manual verification of the tweets in the selected check set. As we assume all wrongly labeled tweets would be corrected in such manual check, our estimations are likely to optimistic.

References

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *International Conference on Web and Social Media*.

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021a. Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):933–942.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021b. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):923–932.

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Mois Aroyo, Maarten Paul Bosma, Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vincent Y. Zhao, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. 2022. Lamda: Language models for dialog applications. In *arXiv*.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards genuinely difficult long context NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16576–16586, Miami, Florida, USA. Association for Computational Linguistics.

Lavanya Gupta, Saket Sharma, and Yiyun Zhao. 2024. Systematic evaluation of long-context LLMs on financial concepts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1163–1175, Miami, Florida, US. Association for Computational Linguistics.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2022. Ruler: What's the real context size of your long-context language models?

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Greg Kamradt. 2023. Llmtest needle in a haystack: Doing simple retrieval from llm models at various context lengths to measure accuracy.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *Preprint*, arXiv:2305.18404.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

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

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

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *Preprint*, arXiv:2408.13006.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, pages 1–8.

Henry Peng Zou, Yue Zhou, Cornelia Caragea, and Doina Caragea. 2023. Crisismatch: Semi-supervised few-shot learning for fine-grained disaster tweet classification. *CoRR*, abs/2310.14627.

A Appendix

A.1 Models

Table 3 contains the information about the 4 LLMs we evaluated and analyzed.

A.2 Datasets

Task 1: humanitarian aid vs. not humanitarian aid

We used data from CrisisBench (Alam et al., 2021b), a consolidated crisis-related social media dataset for humanitarian information processing. We renamed the classes to humanitarian and not humanitarian aid from the original informative vs. not informative classes because these words by themselves were too broad and general. Tweets were annotated as follows(Alam et al., 2021b,a):

- *humanitarian aid*: tweet is useful for humanitarian aid and
- *not humanitarian aid*: tweet is not useful for humanitarian aid.

We sampled from consolidated disaster events from CrisisMMD (Alam et al., 2018) dataset specifically from the following crisis events: California Earthquake, India Floods, Nepal Earthquake, and Vanuatu Cyclone. We randomly sampled 100 tweets for each disaster event.

Table 3: Information of evaluated and analyzed LLMs

Model	Type	Size	Context Length	Source (OpenAI/Huggingface)
GPT-4o-mini	closed	-	128K	gpt-4o-2024-08-06
GPT-4o	closed	-	128K	gpt-4o-mini-2024-07-18
Llama3.1	open	8B	128K	meta-llama/Meta-Llama-3.1-8B-Instruct
Mistral-base	open	7B	32K	mistralai/Mistral-7B-Instruct-v0.3

Task 2: Humanitarian Aid Information Classification

For the humanitarian information classification task, we utilized human-annotated crisis-related tweets from (Alam et al., 2021a). We sampled across four different disaster types: earthquake, hurricane, wildfire and flood. We chose the event with the highest inter-annotator agreement per disaster type based on (Alam et al., 2021a). The original dataset had 11 labels, however, we limited our labels to the 5 that were present in all of our selected crisis events, following (Zou et al., 2023) who also reduced their labels to 7. Originally, we experimented with including the labels: other relevant information and not humanitarian, however, this seemed to be too challenging for the LLM. The humanitarian aid information labels are as follows:

- **Caution and advice**: Reports of warnings issued or lifted, guidance and tips related 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;
- **Injured or dead people**: Reports of injured or dead people due to the disaster;
- **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.;
- **Sympathy and support**: Tweets with prayers, thoughts, and emotional support;

We sampled the test sets of the following crisis events: Mexico Earthquake 2017, Hurricane Harvey 2017, California Wildfires 2017 and Sri Lanka Floods 2017. We randomly sampled 100 tweets for each disaster event.

A.3 Prompts

A.3.1 Classification Prompts

The disaster tweet classification prompts are shown in figures 4 and 5.

```
You will be provided with a tweet. Your task is to
classify the tweet as either "humanitarian aid" or
"not humanitarian aid" based on its content.
Criteria for Classification:
humanitarian aid:
Classify the tweet as "humanitarian aid" if it
contains one or more of the following:
Caution, advice, or warnings (e.g., evacuation
notices, weather alerts).Information about injured,
dead, or affected people. Rescue efforts,
volunteering activities, or donation requests.
Mentions of damage to homes, roads, bridges, or
buildings. References to natural disasters (e.g.,
floods, earthquakes, fires, strong winds).Disaster
area maps or other logistical information.
not humanitarian aid:
Classify the tweet as "not humanitarian aid" if it
does not include any information relevant to
humanitarian assistance or disaster response.
Class Label:
Only assign one of the following two labels. Do not
explain.
humanitarian aid
not humanitarian aid
```

Figure 4: Prompt for Task 1: Humanitarian Aid vs. Not Humanitarian Aid

```
You will be provided a tweet. Based on the tweet's
content, assign one of the following labels related
to humanitarian aid that best fits the information
provided:
Caution and advice: Reports of warnings issued or
lifted, guidance and tips related 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;
Injured or dead people: Reports of people injured or
dead due to the disaster;
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.;
Sympathy and support: Tweets with prayers, thoughts,
and emotional support;
Select only one label, even if multiple labels seem
to apply. Respond with only the label.
Do not add additional information.
Label: <string>
```

Figure 5: Prompt for Task 2: Humanitarian Information Classsification

A.3.2 Check Set Selection Prompts

The prompts for the two strategies of check set selection are in figures 6 and 7.

A.4 Disaster Tweet Classifier Performance

The performance of the LLMs as disaster tweet classifiers are in tables 4 and 5.

A.5 Individual Confidence Elicitation Results

We wanted to know if there is an optimal check set size, compared to the current 20%, from our models by mapping the effective accuracies achieved by the models across changing check set sizes as seen in figure 8.

```
Provide the probability that an AI Assistant's response is correct, as a value between 0.0 and 1.0
for the following task.
    Give only the probability — no words, explanations, or extra commentary whatsoever.
    Respond only with the probability in this format:
    <value between 0.0 and 1.0 only>
```

```
Task: {Classification Task Prompt}
Tweet: {Tweet}
AI Assistant response: {Classification}
```

Figure 6: Prompt for Individual Confidence Elicitation

```
Task Overview:
    You are provided with a list of tweets, each labeled with a classification assigned by an AI Assistant (also an
LLM). Your role is to identify tweets where the assigned classification may not accurately reflect the content,
potentially indicating an error by the AI Assistant.

Instructions:
1. Review the AI Assistant's Classification Prompt: Refer to the prompt used to instruct the AI Assistant on how to
classify the tweets. This prompt outlines the criteria you'll use to assess the accuracy of each classification.
2. Evaluate Classifications: For each tweet, determine if the assigned class aligns with the tweet's content based on
the AI Assistant's classification prompt. Emphasize consistency, especially among tweets with similar themes or
content.
3. Identify Misclassifications: Flag tweets where the assigned class does not match the content according to the AI
Assistant's classification prompt.
4. Select Exactly {COUNT} Tweets: Choose **precisely {COUNT} unique tweets** with classification errors—** no more and
no less**. If you identify more errors than the required count, prioritize tweets that are the most clearly
misclassified.
5. Record Selected Tweets: Include the complete text of each selected tweet verbatim.
6. Use Only Provided Tweets: Choose tweets exclusively from the provided list; do not add, modify, or invent tweets.
7. Avoid Duplicates: Ensure each selected tweet appears only once.

Output Format:
    Your output must include exactly {COUNT} tweets, formatted as a Python list. Do not add any explanation.
    ["<tweet1>", "<tweet2>", ..., "<tweet{COUNT}>"]

    Failure to provide exactly {COUNT} tweets will be considered incorrect output.
```

```
AI Assistant's Classification Prompt:
    {Classification Task Prompt}

Tweet || Class Assigned by AI Assistant:
    {Tweet and Classifications List}
```

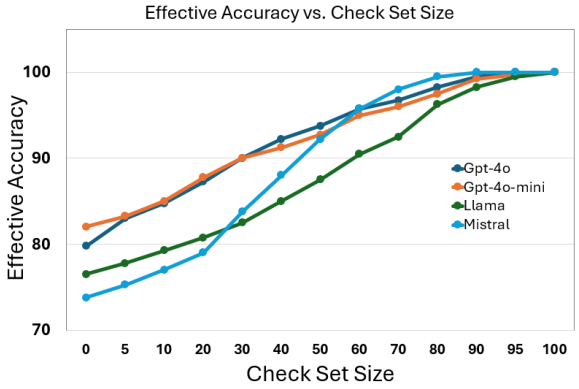
Figure 7: Prompt for Direct Set Selection

Table 4: Performance of LLMs on Task 1: Humanitarian Aid vs. Not Humanitarian Aid measured in Accuracy.

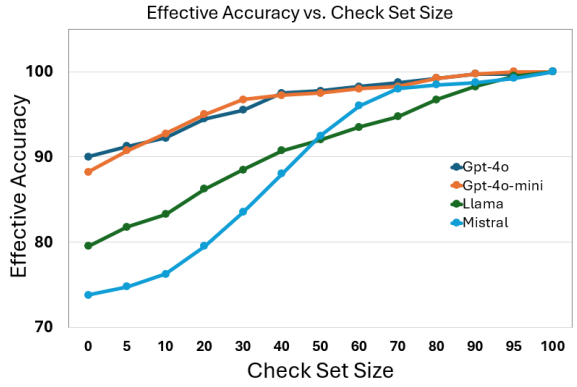
Model	California Earthquake	Vanuatu Cyclone	Nepal Earthquake	India Floods
gpt-4o-mini	0.77	0.82	0.78	0.91
gpt-4o	0.72	0.78	0.77	0.92
llama 3.1	0.62	0.82	0.77	0.85
mistral v.03	0.67	0.73	0.72	0.83
majority class	0.69	0.63	0.50	0.76

Table 5: Performance of LLMs on Task 2: the Humanitarian Aid Information Classification task measured in Accuracy

Model	Mexico Earthquake	Sri Lanka Floods	California Wildfires	Hurricane Harvey
gpt-4o-mini	0.84	0.89	0.94	0.86
gpt-4o	0.89	0.91	0.95	0.85
llama 3.1	0.77	0.84	0.86	0.71
mistral v.03	0.68	0.73	0.84	0.58
majority class	0.35	0.61	0.51	0.23



Task 1: Humanitarian Aid vs. Not Humanitarian Aid



Task 2: Humanitarian Aid Information Classification

Figure 8: Effective Accuracy vs. Check Set Size