# Evaluating Large Language Models for Confidence-based Check Set Selection

**Jane Arleth dela Cruz** **Iris Hendrickx** **Martha Larson**
Center for Language and Speech Technology
Center for Language Studies
Radboud University, Nijmegen, Netherlands
{janearleth.delacruz, iris.hendrickx, martha.larson}@ru.nl

## Abstract

Large Language Models (LLMs) have shown promise in automating high-labor data tasks, but the adoption of LLMs in high-stake scenarios faces two key challenges: their tendency to answer despite uncertainty and their difficulty handling long input contexts robustly. We investigate commonly used off-the-shelf 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* – LLM 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 results reveal that set selection via individual probabilities is more reliable but that direct set confidence merits further investigation. Direct set selection challenges include inconsistent outputs, incorrect check set size, and low inter-annotator agreement. Despite these challenges, our approach improves collaborative disaster tweet classification by outperforming random-sample check set selection, 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 end users 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 challenging to evaluate automatically in high-stakes scenarios, 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 a list of potentially misclassified predictions by the LLM needing review by the end users. While prior research has evaluated the quality of LLM-generated output for escalation to human review, such efforts have typically relied on separate verifier models (Wang et al., 2024; Varshney and Baral, 2023), task-specific fine-tuning (Xin et al., 2021; Chen et al., 2023), or probing the model (Yoshikawa and Okazaki, 2023). In contrast, this study introduces a novel approach in which the check set is directly selected by the off-the-shelf LLM itself.

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,



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 16249–16265

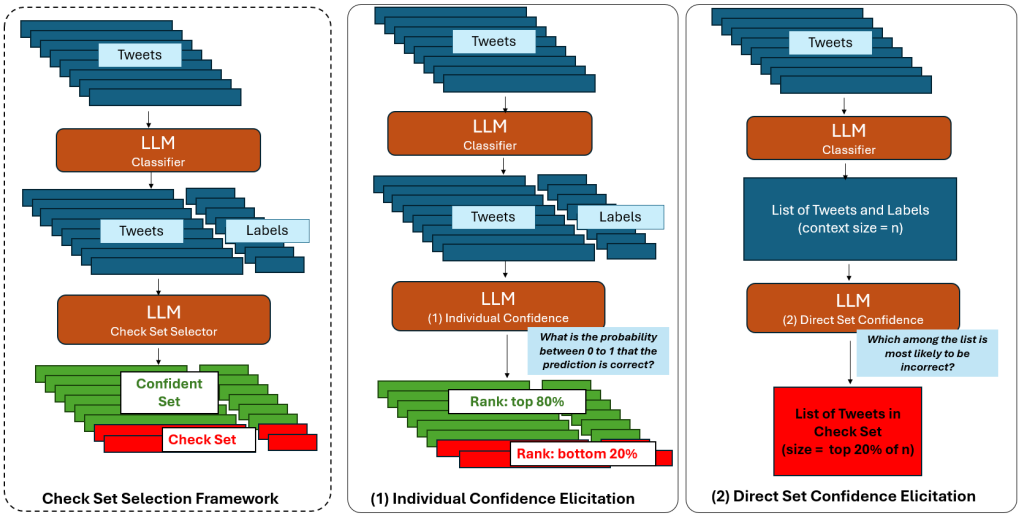



Figure 1: Check Set Selection Framework. Two strategies for check set selection (1) Individual Confidence Elicitation - LLM assesses confidence for each tweet classification individually, requiring more prompts with shorter contexts (2) Direct Set Confidence Elicitation - LLM evaluates confidence for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

enabling human oversight within a budgeted timeframe. LLMs have the potential to assist disaster managers in sifting through massive amounts of online social media data for relevant, critical, and actionable information during disaster events. With the goal of helping disaster managers, we are focused on commonly available LLMs that allow the disaster managers independence from a complex pipeline and the maintenance it implies.

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).

Previous work on selection from long-context lists (Hsieh et al., 2022; Gupta et al., 2024; Levy et al., 2024; Laban et al., 2024), has not required the LLM to identify a subset of specific items within a longer list provided in the prompt and has not explored the influence of the referencing method used for the input. Intuitively, more input data and longer contexts 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 where performance is influenced by the input order and context size (Liu et al., 2024; Hsieh et al., 2022; Gupta et al., 2024).

We ran our experiments using both closed and open-sourced off-the-shelf LLMs: gpt-4o-mini (OpenAI, 2024a), gpt-4o (OpenAI, 2024b), llama 3.1 8B-Instruct (Llama Team, 2024), mistral 7B-Instruct 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.

Our key contributions are as follows:

- We introduce direct set confidence-based check set selection, leveraging fewer prompts with longer context input, and compared it to the individual confidence-based check set selection.
- We analyze the long-context capabilities of LLMs in direct set selection, examining the impact of context length, input order, and referencing methods.

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 demonstrates that the current off-the-shelf LLMs are not sufficiently developed for the direct set selection method. This is evidenced in Section 4 demonstrating the issues in the direct set method such as providing incorrect list sizes, inconsistent outputs across different list-referencing methods, and low inter-annotator agreement. The direct set selection capabilities ought to be further explored as LLMs improve, as the issues pointed out need to be solved, especially when LLMs are to be applied in high stakes scenarios.

## 2 Method

The study investigates 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 two 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 (Zhou et al., 2024) and LLMs struggle with long-context tasks (Liu et al., 2024; Hsieh et al., 2022). We aim to mitigate these problems using check set selection by allowing LLMs to utilize their confidence estimates of their initial predictions to prioritize information needing human review.

**LLM as Disaster Tweet Classifier** We test the performance of LLMs as disaster tweet 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 sixteen (16) different disaster events, Task (1) with 6 disaster events with 500 tweets per event and Task (2) with 10 disaster events with 300 tweets per event. 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 end users and allows us to compare across different check set selection strategies. For cut-off tweets with the same probabilities, we use random selection.

**Evaluating the Reliability of Direct Set Selection** To evaluate the reliability of off-the-shelf LLMs in direct set selection, we conducted a comprehensive analysis of three factors: input context length, list-referencing methods, and input list order. For an LLM to be considered reliable, these factors should not significantly affect its performance.

Our experimental design involved prompting the LLM to identify $k$ tweets with potentially erroneous classification labels from a given list of tweets and classifications provided by an AI assistant. This task requires the LLM to comprehend the initial classification task prompt, access the list of $k$ tweets and classifications, and subsequently select the check set for end-user review. 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 25-tweet context, we partitioned the 100 tweets into four disjoint groups, each prompt selecting five from the list to create the check set size of 20.

Second, we investigate the influence of list-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. Furthermore, as our goal is to allow disaster managers independence from a complex pipeline and optimize resource, the choice of list-referencing method does influence the cost per token (both input and output) and so merits further examination. The four list referencing methods and their rationale are as follows (see Appendix A.4 for examples):

- **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, randomly shuffling the order of tweet classifications in each prompt to assess the influence of list order on selection choices. 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 500 tweets for six different disaster events, i.e., a total of 3000 tweets 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 300 tweets for each of ten different disaster events, i.e., a total of 3000 tweets.

More information about the datasets used is found in Appendix A.2

### 3.2 Models

We chose four of the latest commonly used off-the-shelf LLM's in our experiments. We used gpt-4o-mini (OpenAI, 2024a), gpt-4o (OpenAI, 2024b), llama 3.1-8B-Instruct (Llama Team, 2024), and mistral 7B v0.3-Instruct (Jiang et al., 2023). These models were chosen because they are commonly used by both researchers and the public. 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 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:

$$\%\text{Eff Acc}_D = \frac{(n-c)}{n}\%\text{Acc}_{LLM} + \frac{c}{n}\%\text{Acc}_{Hum}$$

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 consider an LLM 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 list, i.e., there were no hallucinations. We report valid prompt responses by the 100-tweet partitions of a disaster event (our set largest context-size), i.e., one valid response is equivalent to four valid responses of each disjoint group of context size 25 and two valid responses of each disjoint group of context size 50.

**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.

### 3.4 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). We used as our maximum performance metric Mizrahi et al. (2024), accuracy to select the final prompt templates. The exact prompts can be found in the Appendix A.3

**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. Chain-of-thought (CoT) prompting was not explored for set selection as Tian et al., 2023's finding suggest that CoT prompting does not improve verbalized calibration for individual confidence prompting. 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. Specifically, we explored direct set selection and re-ranking the list before selecting the top-$k$ tweets and classifications. From the evaluated prompt strategies, the choice of prompt strategy also influenced the response, so we used the metric, most number of valid prompt responses, to select our final prompts. Figure 7 shows the example direct set selection prompt.

# 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 Tables 3 and 4 measured in accuracy scores at the column Acc. We observed that the closed-source model, gpt-4o-mini performs well in both tasks, achieving accuracy scores of between 74% and 90% for Task 1 and between 86% and 92% 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 Tables 8 and 9 found in Appendix A.5.3. 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 Tables 3 and 4 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 of the same size. We highlighted the instance where the individual confidence check set selection did not outperform random in Tables 3 and 3. We observed that the models, gpt-4o, gpt-4o-mini, and llama individual confidece check set selection outperform random for all the tasks and all the events. Mistral, on the other hand, outperforms random for all except Task (1), Vanuatu cyclone.

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.5.2. 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

**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 number of valid prompt responses LLMs can generate by context size. We observed that the

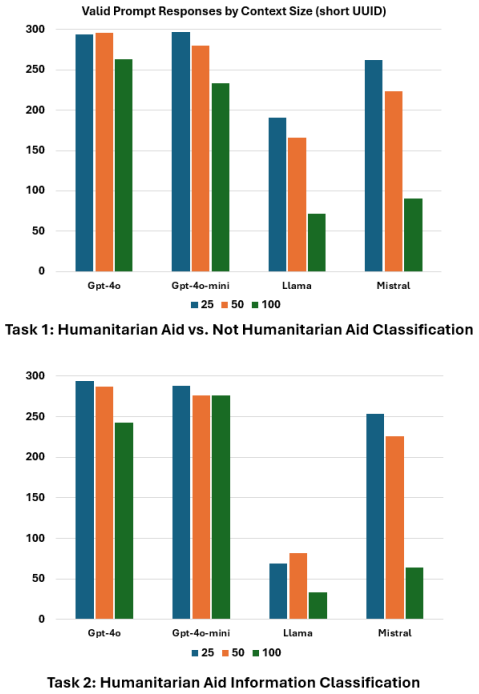



Figure 2: Valid prompt responses by context size using the short UUID referencing method. The values are **raw counts**. One valid prompt response corresponds to one valid check set of size 20 selected from a 100-tweet partition of each disaster event.

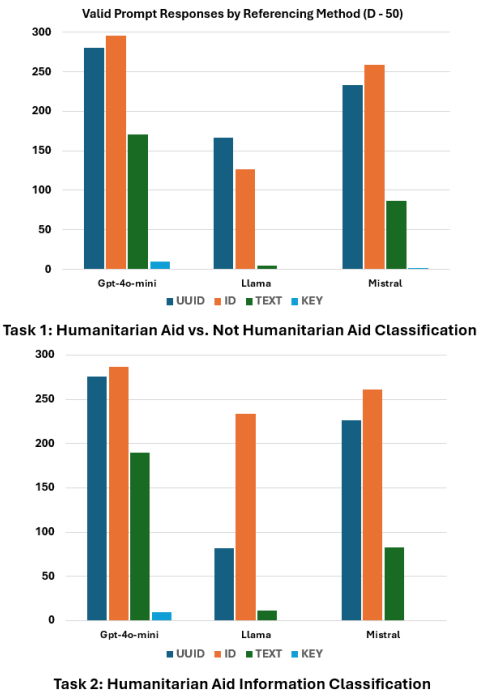



Figure 3: Valid prompt responses by list-referencing method at the 50-tweet context size. The values are **raw counts**. One valid prompt response corresponds to one valid check set of size 20 selected from a 100-tweet partition of a disaster event.

input context size influences LLMs' ability to select a set from a list as seen in Figure 2. We observe that the smallest, 25-tweet context size consistently provides more valid prompt responses accross all models except for llama in Task (2) for the short UUID referencing method.

**The list-referencing method used affects LLM's direct set selection output** Figure 3 shows the number 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 used. We observed that the chosen referencing method affects the number 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 outputs.

**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 in Tables 1 and 2 respectively 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 0.60 for the for Task 1, while Table 2 shows only gpt-4o and llama achieve this. 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 merits further investigation

The effective accuracies from the direct set confidence selection are shown in the columns Eff Acc (D - <context size>) in Tables 3 and 4. 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 depen-

| Task 1: Humanitarian Aid vs. Not Humanitarian Aid | | | | |
|---|---|---|---|---|
| Event | Model | D-100 | D-50 | D-25 |
| California Earthquake | gpt-4o-mini | 0.32 | 0.55 | 0.45 |
| | gpt-4o | 0.30 | 0.33 | 0.35 |
| | llama | 0.25 | 0.25 | 0.20 |
| | mistral | 0.23 | 0.17 | 0.25 |
| Chile Earthquake | gpt-4o-mini | 0.25 | 0.19 | 0.18 |
| | gpt-4o | 0.55 | 0.60 | **0.64** |
| | llama | 0.00 | 0.00 | 0.42 |
| | mistral | 0.09 | 0.11 | 0.16 |
| India Floods | gpt-4o-mini | 0.46 | **0.67** | **0.72** |
| | gpt-4o | **0.69** | **0.72** | 0.71 |
| | llama | 0.25 | 0.00 | 0.00 |
| | mistral | 0.08 | 0.28 | 0.33 |
| Nepal Earthquake | gpt-4o-mini | 0.32 | 0.43 | 0.44 |
| | gpt-4o | 0.35 | 0.45 | 0.48 |
| | llama | 0.00 | 0.07 | 0.00 |
| | mistral | 0.46 | 0.24 | 0.33 |
| Pakistan Earthquake | gpt-4o-mini | 0.24 | 0.37 | 0.38 |
| | gpt-4o | 0.26 | 0.41 | 0.56 |
| | llama | 0.07 | 0.00 | 0.30 |
| | mistral | 0.13 | 0.13 | 0.18 |
| Vanuatu Cyclone | gpt-4o-mini | 0.23 | 0.28 | 0.28 |
| | gpt-4o | 0.43 | **0.64** | **0.66** |
| | llama | 0.07 | 0.00 | 0.33 |
| | mistral | 0.15 | 0.26 | 0.21 |

Table 1: Inter-annotator agreement between the valid prompts. Krippendorf's alpha by context size. **Bold values** indicate high inter-annotator agreement. Short UUID referencing was used.

| Task 2: Humanitarian Information Classification | | | | |
|---|---|---|---|---|
| Event | Model | D-100 | D-50 | D-25 |
| Canada Wildfires | gpt-4o-mini | 0.37 | 0.37 | 0.47 |
| | gpt-4o | 0.50 | 0.47 | **0.67** |
| | llama | 0.00 | 0.06 | 0.43 |
| | mistral | 0.07 | 0.13 | 0.40 |
| Cyclone Idai | gpt-4o-mini | 0.36 | 0.36 | 0.43 |
| | gpt-4o | 0.31 | 0.46 | 0.43 |
| | llama | 0.00 | 0.00 | 0.23 |
| | mistral | 0.07 | 0.15 | 0.32 |
| Greece Wildfires | gpt-4o-mini | 0.18 | 0.18 | 0.29 |
| | gpt-4o | 0.35 | 0.44 | 0.48 |
| | llama | 0.38 | 0.18 | 0.12 |
| | mistral | 0.29 | 0.09 | 0.19 |
| Hurricane Harvey | gpt-4o-mini | 0.41 | 0.41 | 0.45 |
| | gpt-4o | 0.30 | 0.49 | 0.55 |
| | llama | 0.00 | 0.12 | - |
| | mistral | 0.00 | 0.17 | 0.24 |
| Hurricane Maria | gpt-4o-mini | 0.37 | 0.37 | 0.48 |
| | gpt-4o | 0.37 | 0.47 | 0.47 |
| | llama | 0.00 | 0.30 | 0.23 |
| | mistral | 0.19 | 0.10 | 0.20 |
| Hurricane Matthew | gpt-4o-mini | 0.23 | 0.23 | 0.31 |
| | gpt-4o | 0.25 | 0.45 | 0.46 |
| | llama | 0.46 | 0.19 | 0.26 |
| | mistral | 0.22 | 0.10 | 0.30 |
| Italy Earthquake | gpt-4o-mini | 0.28 | 0.28 | 0.33 |
| | gpt-4o | 0.29 | 0.44 | 0.46 |
| | llama | 0.00 | 0.50 | 0.24 |
| | mistral | 0.07 | 0.09 | 0.21 |
| Maryland Floods | gpt-4o-mini | 0.32 | 0.32 | 0.47 |
| | gpt-4o | 0.16 | 0.34 | 0.46 |
| | llama | **1.00** | 0.16 | 0.32 |
| | mistral | 0.10 | 0.15 | 0.14 |
| Mexico Earthquake | gpt-4o-mini | 0.35 | 0.35 | 0.38 |
| | gpt-4o | 0.47 | 0.53 | 0.48 |
| | llama | 0.00 | 0.50 | 0.22 |
| | mistral | 0.34 | 0.12 | 0.26 |
| Sri Lanka Floods | gpt-4o-mini | 0.55 | 0.55 | 0.58 |
| | gpt-4o | **0.67** | 0.55 | 0.59 |
| | llama | 0.07 | 0.18 | 0.33 |
| | mistral | 0.05 | 0.20 | 0.26 |

Table 2: Inter-annotator agreement between the valid prompts. Krippendorf's alpha by context size. **Bold values** indicate high inter-annotator agreement. Short UUID referencing was used.

dent on the luck of the misclassified tweets being evenly distributed across subgroups. When compared with the effective accuracies using random check set, only check set selection using gpt-4o outperforms random across all tasks and context sizes, while gpt-4o-mini, llama and 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 issues across input context length, list-referencing method and input list order. Furthermore, as observed in Tables 3 and 4, only individual confidence outperform random consistently across tasks and events. Full tables of effective accuracies, that also include (*Eff Acc (Random)*) – effective accuracies for random check set and (*Eff Acc (Max)*) – maximum possible effective accuracies, can be found in Tables 8 and 9 in Appendix A.5.3.

## 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) leads to different check set selections and can even return invalid responses. Invalid responses include providing incorrect number of items from the output list (both more or less than asked), repeating items in the output, editing short-UUID's where characters can be replaced in items, and editing output full-text tweets to correct the grammar. This observation holds across different input context sizes. We recommend evaluating LLMs with multiple prompts always as we have observed that this issue is under reported.

We performed a sanity check on the check sets selected by the LLMs. Upon manual evaluation of a subset of disaster events for Task 2, we observed that they were commonly tweets that contain multiple information, meaning they can be classified into more than one category.

We did a quick exploration on inference time vs. Effective Accuracy across our two strategies. The disaster manager wants to keep the use of the model cheap, which is probably related to computational/inference time, but more practically it is related to output tokens. Here, we assume that direct set selection has a clear advantage. In terms of inference time a rough estimate on a 100-tweet sample for Llama to select a checks size of 20, for example takes 18.4 seconds inference time on individual confidence (100 prompts), 3 seconds on direct set (D-100), 3.33 seconds on D-50, and 3.65

| Task 1: Humanitarian Aid vs. Not Humanitarian Aid | | | | | | |
|---|---|---|---|---|---|---|
| Event | Model | Acc | Eff Acc (I) | Eff Acc (D-100) | Eff Acc (D-50) | Eff Acc (D-25) |
| California Earthquake | gpt-4o-mini | 0.74 | 0.80 | 0.86 | 0.86 | 0.78 |
| | gpt-4o | 0.67 | 0.77 | 0.82 | 0.81 | 0.79 |
| | llama | 0.73 | 0.79 | 0.71 | 0.73 | 0.71 |
| | mistral | 0.54 | 0.67 | 0.63 | 0.67 | 0.67 |
| Chile Earthquake | gpt-4o-mini | 0.82 | 0.92 | 0.87 | 0.91 | 0.83 |
| | gpt-4o | 0.73 | 0.85 | 0.91 | 0.91 | 0.90 |
| | llama | 0.73 | 0.80 | 0.78 | 0.78 | 0.76 |
| | mistral | 0.65 | 0.73 | 0.73 | 0.75 | 0.76 |
| India Floods | gpt-4o-mini | 0.90 | 0.95 | 0.95 | 0.97 | 0.96 |
| | gpt-4o | 0.87 | 0.94 | 0.97 | 0.97 | 0.97 |
| | llama | 0.64 | 0.71 | 0.81 | 0.79 | 0.81 |
| | mistral | 0.80 | 0.89 | 0.84 | 0.89 | 0.91 |
| Nepal Earthquake | gpt-4o-mini | 0.82 | 0.90 | 0.89 | 0.90 | 0.87 |
| | gpt-4o | 0.74 | 0.86 | 0.86 | 0.88 | 0.88 |
| | llama | 0.75 | 0.85 | 0.81 | 0.79 | 0.79 |
| | mistral | 0.65 | 0.74 | 0.73 | 0.78 | 0.78 |
| Pakistan Earthquake | gpt-4o-mini | 0.81 | 0.89 | 0.86 | 0.86 | 0.85 |
| | gpt-4o | 0.66 | 0.87 | 0.87 | 0.87 | 0.87 |
| | llama | 0.74 | 0.79 | 0.78 | 0.79 | 0.78 |
| | mistral | 0.67 | 0.75 | 0.73 | 0.78 | 0.78 |
| Vanuatu Cyclone | gpt-4o-mini | 0.87 | 0.94 | 0.91 | 0.91 | 0.88 |
| | gpt-4o | 0.76 | 0.92 | 0.93 | 0.94 | 0.94 |
| | llama | 0.79 | 0.85 | 0.83 | 0.83 | 0.82 |
| | mistral | 0.82 | 0.84 | 0.85 | 0.87 | 0.87 |

Table 3: Effective Accuracies of the Check Set Selection Strategies. Eff Acc (I) is for the individual confidence and Eff Acc (D) is for direct set confidence and the number indicates the context length size. The highlight indicates when the Eff Acc does not outperform random. short UUID-referencing was used. Full table can be found in Table 8.

| Task 2: Humanitarian Aid Information Classification | | | | | | |
|---|---|---|---|---|---|---|
| Event | Model | Acc | Eff Acc (I) | Eff Acc (D-100) | Eff Acc (D-50) | Eff Acc (D-25) |
| Canada Wildfires | gpt-4o-mini | 0.92 | 0.97 | 0.95 | 0.95 | 0.96 |
| | gpt-4o | 0.92 | 0.99 | 0.98 | 0.98 | 0.98 |
| | llama | 0.86 | 0.92 | 0.90 | 0.88 | 0.90 |
| | mistral | 0.86 | 0.94 | 0.87 | 0.89 | 0.90 |
| Cyclone Idai | gpt-4o-mini | 0.87 | 0.94 | 0.90 | 0.91 | 0.92 |
| | gpt-4o | 0.89 | 0.96 | 0.92 | 0.95 | 0.94 |
| | llama | 0.80 | 0.88 | 0.84 | 0.83 | 0.85 |
| | mistral | 0.71 | 0.82 | 0.76 | 0.78 | 0.77 |
| Greece Wildfires | gpt-4o-mini | 0.93 | 0.96 | 0.95 | 0.95 | 0.94 |
| | gpt-4o | 0.92 | 0.97 | 0.95 | 0.97 | 0.96 |
| | llama | 0.81 | 0.85 | 0.85 | 0.84 | 0.84 |
| | mistral | 0.58 | 0.66 | 0.66 | 0.67 | 0.64 |
| Hurricane Harvey | gpt-4o-mini | 0.86 | 0.94 | 0.89 | 0.88 | 0.90 |
| | gpt-4o | 0.89 | 0.95 | 0.93 | 0.94 | 0.94 |
| | llama | 0.75 | 0.85 | 0.80 | 0.79 | 0.80 |
| | mistral | 0.64 | 0.79 | 0.70 | 0.70 | 0.71 |
| Hurricane Maria | gpt-4o-mini | 0.88 | 0.95 | 0.93 | 0.93 | 0.94 |
| | gpt-4o | 0.90 | 0.97 | 0.96 | 0.96 | 0.95 |
| | llama | 0.79 | 0.84 | 0.82 | 0.82 | 0.83 |
| | mistral | 0.76 | 0.88 | 0.83 | 0.83 | 0.80 |
| Hurricane Matthew | gpt-4o-mini | 0.88 | 0.95 | 0.91 | 0.93 | 0.92 |
| | gpt-4o | 0.91 | 0.97 | 0.96 | 0.96 | 0.96 |
| | llama | 0.77 | 0.84 | 0.82 | 0.81 | 0.82 |
| | mistral | 0.65 | 0.72 | 0.72 | 0.73 | 0.71 |
| Italy Earthquake | gpt-4o-mini | 0.92 | 0.94 | 0.93 | 0.94 | 0.94 |
| | gpt-4o | 0.92 | 0.97 | 0.96 | 0.96 | 0.96 |
| | llama | 0.86 | 0.89 | 0.88 | 0.89 | 0.88 |
| | mistral | 0.66 | 0.74 | 0.72 | 0.71 | 0.71 |
| Maryland Floods | gpt-4o-mini | 0.88 | 0.92 | 0.90 | 0.91 | 0.91 |
| | gpt-4o | 0.89 | 0.93 | 0.93 | 0.93 | 0.94 |
| | llama | 0.77 | 0.86 | 0.80 | 0.82 | 0.82 |
| | mistral | 0.62 | 0.75 | 0.68 | 0.70 | 0.69 |
| Mexico Earthquake | gpt-4o-mini | 0.92 | 0.95 | 0.94 | 0.95 | 0.96 |
| | gpt-4o | 0.91 | 0.96 | 0.95 | 0.97 | 0.96 |
| | llama | 0.85 | 0.89 | 0.88 | 0.89 | 0.89 |
| | mistral | 0.78 | 0.89 | 0.81 | 0.81 | 0.83 |
| Sri Lanka Floods | gpt-4o-mini | 0.92 | 0.97 | 0.94 | 0.96 | 0.96 |
| | gpt-4o | 0.94 | 0.98 | 0.98 | 0.98 | 0.98 |
| | llama | 0.90 | 0.93 | 0.91 | 0.93 | 0.93 |
| | mistral | 0.82 | 0.92 | 0.85 | 0.86 | 0.87 |

Table 4: Effective Accuracies of the Check Set Selection Strategies. Eff Acc (I) is for the individual confidence and Eff Acc (D) is for direct set confidence and the number indicates the context length size. The highlight indicates when the Eff Acc does not outperform random. short UUID-referencing was used. Full table can be found in Table 9.

seconds on D-25. We recommend a more thorough investigation on these factors as LLMs improve.

## 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, so called 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) more accurately than their raw conditional probabilities suggest. For our individual-based check set selection, we used verbalized numerical confidence.

**Selective Prediction for LLMs.** Prior research on selective prediction and escalation for human review in LLMs generally follows three main strategies. First, separate verifier models are trained as external classifiers to identify uncertain or potentially incorrect outputs without modifying the base LLM, as explored in human-LLM collaborative annotation frameworks (Wang et al., 2024; Varshney and Baral, 2023). Ma et al., 2023's filter-then-rerank paradigm employs a separate small language model as a verifier model for the LLM Second, task-specific fine-tuning adapts the LLM itself to better estimate uncertainty by incorporating error regularization or self-evaluation during training (Chen et al., 2023; Xin et al., 2021; Lin and Ma, 2024). Third, model-probing techniques analyze the LLM's internal signals, for example, Selective-LAMA uses token-level confidence thresholds to filter dubious predictions (Yoshikawa and Okazaki, 2023). Unlike these strategies, which require external verifiers, task-specific training, or manual probing – our approach directly leverages the off-the-shelf LLM's own confidence estimates to curate

check sets for human review, simplifying deployment and broadening applicability

**LLM performance on long-context input text**. For the direct-set check set selection we propose, we explored long-context prompts, which are previously studied in, e.g., Hsieh et al. (2022); Shaham et al. (2023); Levy et al. (2024); Laban 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 most common task that papers evaluate on is the needle-in-a-haystack (NIAH) task, where the LLMs are tasked to retrieve single points (the "needle") in a long input context (the "haystack") and asking the LLM to retrieve it given a related question (Kamradt, 2023) and not multiple needles. Hsieh et al. (2022) expands the NIAH task with a comprehensive evaluation of long-context LLMs by creating a new synthetic benchmark revealing that almost all models exhibit large performance drops as context increases. Most papers evaluate LLM performance on synthetic datasets or existing benchmarks (Hsieh et al., 2022; Shaham et al., 2023; Levy et al., 2024; Laban et al., 2024). The study by Gupta et al. (2024) differs by evaluating LLMs in a real-world financial dataset, however, evaluated only the gpt-4 suite of LLMs in solving tasks, as a function of factors such as context length, task difficulty, and position of needle. Our study on the other hand, evaluates both off-the-shelf closed and open-sourced LLMs and considers list-referencing factors in addition to the context length and input list order on real-world crisis-related tweets.

## 7 Conclusion

In this paper, we investigate the ability of LLMs 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 for social media monitoring 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 the impact of context length, input order and referencing methods for direct set selection. Our results show that LLMs struggle in direct set selection as they cannot consistently provide valid prompt responses, being influenced by all the three factors mentioned. 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 and evaluated further as LLMs continue to improve. Despite these challenges, our approach improves collaborative disaster tweet classification, demonstrating the potential of human-LLM collaboration. Such collaboration is crucial for high stake scenarios where we want the end-user in control of the final decisions.

## 8 Limitations

We only evaluated four commonly used off-the-shelf 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 too optimistic.

## 9 Ethical Considerations

The datasets used in this paper were from publicly available datasets (Alam et al., 2021b,a) which were collected tweets from X (previously, Twitter) using the platform's streaming API in line with its terms of service.

Our work aspires ultimately to support disaster management in high-stakes scenarios. As such, a potential risk is that readers misinterpret the readiness of the technology for use by disaster managers, and move either too quickly to uptake without guarantees of reliability or pre-maturely abandon the

type of solutions we study. We have attempted to address this point by stating clearly our **negative result** (i.e., LLMs struggle with long-context set selection) and stating that we find human-LLM collaborations may still hold future potential.

## Acknowledgments

This publication is part of the project 'Indeep: Interpreting Deep Learning Models for Text and Sound' with project number NWA.1292.19.399, which is partly financed by the Dutch Research Council (NWO).

## 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.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. 2023. Adaptation with self-evaluation to improve selective prediction in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213, Singapore. Association for Computational Linguistics.

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. In *Neural Conversational AI Workshop at ICML 2023*.

Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages

9885–9903, Miami, Florida, USA. Association for Computational Linguistics.

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*.

Yu-Xiang Lin and Wei-Yun Ma. 2024. Generating attractive and authentic copywriting from customer reviews. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4629–4642, Mexico City, Mexico. Association for Computational Linguistics.

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.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

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.

Neeraj Varshney and Chitta Baral. 2023. Post-abstention: Towards reliably re-attempting the abstained instances in QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 967–982, Toronto, Canada. Association for Computational Linguistics.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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. In *Building Trust in LLMs and LLM Applications workshop at ICLR 2025*.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

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).

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*,

pages 2017–2028, Dubrovnik, Croatia. Association for Computational Linguistics.

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 5 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: Pakistan Earthquake 2013, California Earthquake 2014, Chile Earthquake 2014, India Floods 2014, Nepal Earthquake 2014, and Vanuatu Cyclone 2014. We randomly sampled 500 tweets for each disaster event.

**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: Canada Wildfires 2016, Cyclone Idai 2019, Greece Wildfires 2018, Mexico Earthquake 2017, Hurricane Matthew 2016, Hurricane Harvey 2017, Hurricane Maria 2017, Italy Earthquake 2016, Maryland Floods 2018, and Sri Lanka Floods 2017. We randomly sampled 300 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.

### 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 Output Examples by List-referencing Method

Below are output examples of valid responses by list-referencing method at 50-tweet context size.

**Numerical ID:**

```
[366, 191, 233, 356, 74, 149, 80, 242,
```

Table 5: 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 |
| llama 3.1 8B - Instruct | open | 8B | 128K | meta-llama/Meta-llama-3.1-8B-Instruct |
| mistral 7B - Instruct v0.3 | open | 7B | 32K | mistralai/mistral-7B-Instruct-v0.3 |

```
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


282, 301, 317, 290, 175, 349, 10, 1, 2, 55, 7, 14]

**short UUID**:

['d8d26064', '88ef4c41', '9cb96943', '41bb8105', '785935c5', 'ea8dfa5b', '4eeff954', '60df1292', 'b6f5170d', '2b577377']

**Key Word**:

["distributing commodities", "donate to help", "Hurricane Maria Disaster Recovery", "donate for hurricane relief", "devastated by Hurricane Maria", "damaged Puerto Rico", "death toll climbs", "damaged Arecibo radio telescope", "ruined homes and infrastructure", "donations with what you can"]

**Full Text**:

[ "80 hours! ! GOD!! https://t.co/sNetLbIskQ","RT @USER: .@USER Lives may have been saved if Nepal govt prepared people instead of funding animal sacrific Gadhimai ht","RT @USER: 38,000 Nepal youth in Indian Army Gorka Rifles. Over 1.25 lakh veterans. The family will come together in thi hour of c", "Big day for nepal people", "Pulitzer Prize winning Jim Morin's cartoon on NepalQuake NepalEarthquake URL", "but our farmers issues are gone unnoticed URL", "@USER You are amazing. URL", "Economic Impact Of Nepal Quake Likely To Be Massive: One estimate puts the reconstruction at more than $5 bill.. URL", "@USER Huh. I guess all those Christian missions to Nepal are to protect 7-11's Himlayan locations.", "12 Things Indians Can Do To HelpåÊNepal URL" ]


** edited Full Text responses by anonymizing users and URL's.

### A.5 Supplementary Results

#### A.5.1 Disaster Tweet Classifier Performance

The performance of the LLMs as disaster tweet classifiers are in Tables 6 and 7.

#### A.5.2 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}
```

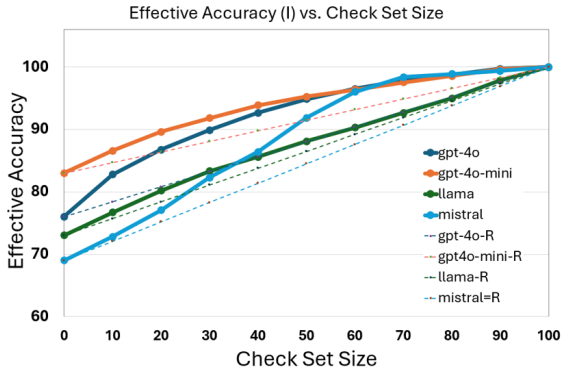
Figure 7: Prompt for Direct Set Selection

### A.5.3 Effective Accuracies Full Tables

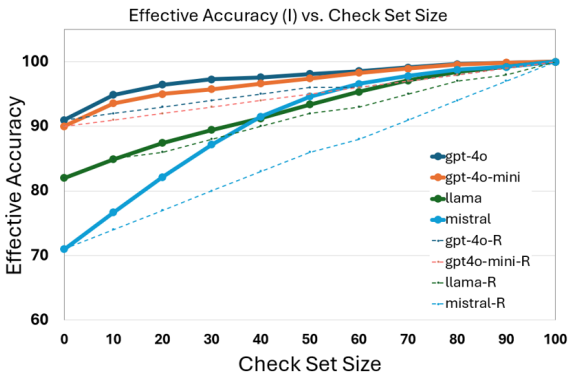
Tables 8 and 9 show the full tables effective accuracies of all the check set selection strategies.

Table 6: Performance of LLMs on Task 1: Humanitarian Aid vs. Not Humanitarian Aid measured in Accuracy.

| Event | Model | Accuracy |
|---|---|---|
| California Earthquake | majority class | 0.84 |
| | gpt-4o-mini | 0.74 |
| | gpt-4o | 0.67 |
| | llama | 0.73 |
| | mistral | 0.54 |
| Chile Earthquake | majority class | 0.60 |
| | gpt-4o-mini | 0.82 |
| | gpt-4o | 0.73 |
| | llama | 0.73 |
| | mistral | 0.65 |
| India Floods | majority class | 0.86 |
| | gpt-4o-mini | 0.90 |
| | gpt-4o | 0.87 |
| | llama | 0.64 |
| | mistral | 0.80 |
| Nepal Earthquake | majority class | 0.72 |
| | gpt-4o-mini | 0.82 |
| | gpt-4o | 0.74 |
| | llama | 0.75 |
| | mistral | 0.65 |
| Pakistan Earthquake | majority class | 0.61 |
| | gpt-4o-mini | 0.81 |
| | gpt-4o | 0.66 |
| | llama | 0.74 |
| | mistral | 0.67 |
| Vanuatu Cyclone | majority class | 0.50 |
| | gpt-4o-mini | 0.87 |
| | gpt-4o | 0.76 |
| | llama | 0.79 |
| | mistral | 0.82 |




**Task 1: Humanitarian Aid vs. Not Humanitarian Aid**




**Task 2: Humanitarian Aid Information Classification**

Figure 8: Effective Accuracy (Individual Confidence) vs. Check Set Size. The broken lines represent the Effective Accuracies for the random check set selection.

Table 7: Performance of LLMs on Task 2: the Humanitarian Aid Information Classification task measured in Accuracy

| Event | Model | Accuracy |
|---|---|---|
| Canada Wildfires | majority class | 0.67 |
| | gpt-4o-mini | 0.92 |
| | gpt-4o | 0.92 |
| | llama | 0.86 |
| | mistral | 0.86 |
| Cyclone Idai | majority class | 0.52 |
| | gpt-4o-mini | 0.87 |
| | gpt-4o | 0.89 |
| | llama | 0.80 |
| | mistral | 0.71 |
| Greece Wildfires | majority class | 0.40 |
| | gpt-4o-mini | 0.93 |
| | gpt-4o | 0.92 |
| | llama | 0.81 |
| | mistral | 0.58 |
| Hurricane Harvey | majority class | 0.45 |
| | gpt-4o-mini | 0.86 |
| | gpt-4o | 0.89 |
| | llama | 0.75 |
| | mistral | 0.64 |
| Hurricane Maria | majority class | 0.43 |
| | gpt-4o-mini | 0.88 |
| | gpt-4o | 0.9 |
| | llama | 0.79 |
| | mistral | 0.76 |
| Hurricane Matthew | majority class | 0.33 |
| | gpt-4o-mini | 0.88 |
| | gpt-4o | 0.91 |
| | llama | 0.77 |
| | mistral | 0.65 |
| Italy Earthquake | majority class | 0.52 |
| | gpt-4o-mini | 0.92 |
| | gpt-4o | 0.92 |
| | llama | 0.86 |
| | mistral | 0.66 |
| Maryland Floods | majority class | 0.29 |
| | gpt-4o-mini | 0.88 |
| | gpt-4o | 0.89 |
| | llama | 0.77 |
| | mistral | 0.62 |
| Mexico Earthquake | majority class | 0.52 |
| | gpt-4o-mini | 0.92 |
| | gpt-4o | 0.91 |
| | llama | 0.85 |
| | mistral | 0.78 |
| Sri Lanka Floods | majority class | 0.70 |
| | gpt-4o-mini | 0.92 |
| | gpt-4o | 0.94 |
| | llama | 0.9 |
| | mistral | 0.82 |

**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 | 0.74 | *0.79* | *0.94* | 0.80 | 0.86 | 0.86 | 0.78 |
| | gpt-4o | 0.67 | *0.74* | *0.87* | 0.77 | 0.82 | 0.81 | 0.79 |
| | llama | 0.73 | *0.78* | *0.93* | 0.79 | 0.71 | 0.73 | 0.71 |
| | mistral | 0.54 | *0.63* | *0.74* | 0.67 | 0.63 | 0.67 | 0.67 |
| Chile Earthquake | gpt-4o-mini | 0.82 | *0.86* | *1.00* | 0.92 | 0.87 | 0.91 | 0.83 |
| | gpt-4o | 0.73 | *0.78* | *0.93* | 0.85 | 0.91 | 0.91 | 0.90 |
| | llama | 0.73 | *0.78* | *0.93* | 0.80 | 0.78 | 0.78 | 0.76 |
| | mistral | 0.65 | *0.72* | *0.85* | 0.73 | 0.73 | 0.75 | 0.76 |
| India Floods | gpt-4o-mini | 0.90 | *0.92* | *1.00* | 0.95 | 0.95 | 0.97 | 0.96 |
| | gpt-4o | 0.87 | *0.90* | *1.00* | 0.94 | 0.97 | 0.97 | 0.97 |
| | llama | 0.64 | *0.71* | *0.84* | 0.71 | 0.81 | 0.79 | 0.81 |
| | mistral | 0.80 | *0.84* | *1.00* | 0.89 | 0.84 | 0.89 | 0.91 |
| Nepal Earthquake | gpt-4o-mini | 0.82 | *0.86* | *1.00* | 0.90 | 0.89 | 0.90 | 0.87 |
| | gpt-4o | 0.74 | *0.79* | *0.94* | 0.86 | 0.86 | 0.88 | 0.88 |
| | llama | 0.75 | *0.80* | *0.95* | 0.85 | 0.81 | 0.79 | 0.79 |
| | mistral | 0.65 | *0.72* | *0.85* | 0.74 | 0.73 | 0.78 | 0.78 |
| Pakistan Earthquake | gpt-4o-mini | 0.81 | *0.85* | *1.00* | 0.89 | 0.86 | 0.86 | 0.85 |
| | gpt-4o | 0.66 | *0.73* | *0.86* | 0.87 | 0.87 | 0.87 | 0.87 |
| | llama | 0.74 | *0.79* | *0.94* | 0.79 | 0.78 | 0.79 | 0.78 |
| | mistral | 0.67 | *0.74* | *0.87* | 0.75 | 0.73 | 0.78 | 0.78 |
| Vanuatu Cyclone | gpt-4o-mini | 0.87 | *0.90* | *1.00* | 0.94 | 0.91 | 0.91 | 0.88 |
| | gpt-4o | 0.76 | *0.81* | *0.96* | 0.92 | 0.93 | 0.94 | 0.94 |
| | llama | 0.79 | *0.83* | *0.99* | 0.85 | 0.83 | 0.83 | 0.82 |
| | mistral | 0.82 | *0.86* | *1.00* | 0.84 | 0.85 | 0.87 | 0.87 |

Table 8: 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

**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) |
|---|---|---|---|---|---|---|---|---|
| Canada Wildfires | gpt-4o-mini | 0.92 | *0.94* | *1.00* | 0.97 | 0.95 | 0.95 | 0.96 |
| | gpt-4o | 0.92 | *0.94* | *1.00* | 0.99 | 0.98 | 0.98 | 0.98 |
| | llama | 0.86 | *0.89* | *1.00* | 0.92 | 0.90 | 0.88 | 0.90 |
| | mistral | 0.86 | *0.89* | *1.00* | 0.94 | 0.87 | 0.89 | 0.90 |
| Idai | gpt-4o-mini | 0.87 | *0.90* | *1.00* | 0.94 | 0.90 | 0.91 | 0.92 |
| | gpt-4o | 0.89 | *0.91* | *1.00* | 0.96 | 0.92 | 0.95 | 0.94 |
| | llama | 0.80 | *0.84* | *1.00* | 0.88 | 0.84 | 0.83 | 0.85 |
| | mistral | 0.71 | *0.77* | *0.91* | 0.82 | 0.76 | 0.78 | 0.77 |
| Greece Wildfires | gpt-4o-mini | 0.93 | *0.94* | *1.00* | 0.96 | 0.95 | 0.95 | 0.94 |
| | gpt-4o | 0.92 | *0.94* | *1.00* | 0.97 | 0.95 | 0.97 | 0.96 |
| | llama | 0.81 | *0.85* | *1.00* | 0.85 | 0.85 | 0.84 | 0.84 |
| | mistral | 0.58 | 0.66 | *0.78* | *0.73* | 0.66 | 0.67 | 0.64 |
| Hurricane Harvey | gpt-4o-mini | 0.86 | *0.89* | *1.00* | 0.94 | 0.89 | 0.88 | 0.90 |
| | gpt-4o | 0.89 | *0.91* | *1.00* | 0.95 | 0.93 | 0.94 | 0.94 |
| | llama | 0.75 | *0.80* | *0.95* | 0.85 | 0.80 | 0.79 | 0.80 |
| | mistral | 0.64 | *0.71* | *0.84* | 0.79 | 0.70 | 0.70 | 0.71 |
| Hurricane Maria | gpt-4o-mini | 0.88 | *0.90* | *1.00* | 0.95 | 0.93 | 0.93 | 0.94 |
| | gpt-4o | 0.90 | *0.92* | *1.00* | 0.97 | 0.96 | 0.96 | 0.95 |
| | llama | 0.79 | *0.83* | *0.99* | 0.84 | 0.82 | 0.82 | 0.83 |
| | mistral | 0.76 | *0.81* | *0.96* | 0.88 | 0.83 | 0.83 | 0.80 |
| Hurricane Matthew | gpt-4o-mini | 0.88 | *0.90* | *1.00* | 0.95 | 0.91 | 0.93 | 0.92 |
| | gpt-4o | 0.91 | *0.93* | *1.00* | 0.97 | 0.96 | 0.96 | 0.96 |
| | llama | 0.77 | *0.82* | *0.97* | 0.84 | 0.82 | 0.81 | 0.82 |
| | mistral | 0.65 | 0.72 | *0.85* | *0.75* | 0.72 | 0.73 | 0.71 |
| Italy Earthquake | gpt-4o-mini | 0.92 | *0.94* | *1.00* | 0.94 | 0.93 | 0.94 | 0.94 |
| | gpt-4o | 0.92 | *0.94* | *1.00* | 0.97 | 0.96 | 0.96 | 0.96 |
| | llama | 0.86 | *0.89* | *1.00* | 0.89 | 0.88 | 0.89 | 0.88 |
| | mistral | 0.66 | *0.73* | *0.86* | 0.74 | 0.72 | 0.71 | 0.71 |
| Italy Earthquake | gpt-4o-mini | 0.88 | *0.90* | *1.00* | 0.92 | 0.90 | 0.91 | 0.91 |
| | gpt-4o | 0.89 | *0.91* | *1.00* | 0.93 | 0.93 | 0.93 | 0.94 |
| | llama | 0.77 | *0.82* | *0.97* | 0.86 | 0.80 | 0.82 | 0.82 |
| | mistral | 0.62 | *0.70* | *0.82* | 0.75 | 0.68 | 0.70 | 0.69 |
| Maryland Floods | gpt-4o-mini | 0.92 | *0.94* | *1.00* | 0.95 | 0.94 | 0.95 | 0.96 |
| | gpt-4o | 0.91 | *0.93* | *1.00* | 0.96 | 0.95 | 0.97 | 0.96 |
| | llama | 0.85 | *0.88* | *1.00* | 0.89 | 0.88 | 0.89 | 0.89 |
| | mistral | 0.78 | *0.82* | *0.98* | 0.89 | 0.81 | 0.81 | 0.83 |
| Sri Lanka Floods | gpt-4o-mini | 0.92 | *0.94* | *1.00* | 0.97 | 0.94 | 0.96 | 0.96 |
| | gpt-4o | 0.94 | *0.95* | *1.00* | 0.98 | 0.98 | 0.98 | 0.98 |
| | llama | 0.90 | *0.92* | *1.00* | 0.93 | 0.91 | 0.93 | 0.93 |
| | mistral | 0.82 | *0.86* | *1.00* | 0.92 | 0.85 | 0.86 | 0.87 |

Table 9: 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