EasyChair Preprint

№ 15820

Advancing AI Incidents Classification: Leveraging LLMs with Strategic Prompting

Yian Chen, Lana Do, Liheng Yi, Ricardo Baeza-Yates and John A. Guerra-Gomez

February 11, 2025

Advancing AI Incidents Classification: Leveraging LLMs with Strategic Prompting

Yian Chen, Lana Do, Liheng Yi, Ricardo Baeza-Yates[0000−0003−3208−9778], and John A. Guerra-Gomez[0000−0001−7943−0000]

Northeastern University, Oakland, CA 94613, USA
{chen.yian1, do.ng, yi.lih, r.baeza-yates, jguerra}@northeastern.edu

Abstract. This study examines the efficacy of large language models (LLMs), particularly GPT-4, in classifying AI incident reports documented in the AI Incidents Database (AIID), with the goal of enhancing our understanding and management of AI-related harm. The data of incident reports are all on news events that detail specific incidents related to AI technology that have resulted in harmful effect on our society. We explore the use of different prompting techniques on GPT-4 and assess the classification results of those incidents by subjective and objective evaluations. This work lays the groundwork for a comprehensive, automated classification framework for AI incident reporting, balancing LLM capabilities with the intricacies inherent in human judgment.

Keywords: Large Language Models · LLM Classification · Prompt Engineering · Responsible AI.

1 Introduction

The proliferation of artificial intelligence (AI) technologies has introduced unprecedented opportunities for innovation across diverse sectors. However, with these advancements come emerging concerns about AI failures. Instances of AI causing harm and sparking controversies underscore the critical need for effective classification frameworks to systematically categorize these incidents.

As part of a larger study of Responsible AI [1], this study primarily focuses on the development of a robust classification methodology to categorize a vast volume of technology-related incidents sourced from the AI Incidents Database(AIID)[1]. The AIID serves as a repository of news articles documenting instances where AI technologies have caused harmful outcomes in the real-world.

Our objective is to streamline the process of the incidents classification for data visualization. However, the sheer volume of incidents—exceeding 600 cases—renders manual classification impractical. To address this challenge, we propose leveraging large language models (LLMs) to to automate the classification process.

[1] https://incidentdatabase.ai/

Pre-trained transformer language models, such as GPT-3 and GPT-4, demonstrate great in-context learning capabilities through natural language task description [2]. Those LLMs exhibit breakthrough performances in various downstream tasks such as text-interpretation through semantic processing abilities [4] and medical diagnoses or theorem proving through its reasoning capabilities by general deduction [13]. We harness the power of these Artificial Intelligence Generated Content (AIGC) tools for our classification task by providing the data from AIID to the LLM and use various prompting techniques to refine the results. In our use-case, it is critical to ensure that the automated classifications are not only well-supported by the data given, but also closely aligned with human judgment and standards. This necessitates a validation process to ascertain the reliability and trustworthiness of the classification outcomes against established human benchmarks.

By utilizing LLM technology, we aim to develop a scalable and efficient approach to classify these AI-related incidents. Our efforts are geared towards providing insights into the prevalence, trends, and implications of AI-related harm, thereby contributing to the advancement of responsible AI researches.

AI engineering is an emerging field, and consequently, AI incidents documentation practices are still evolving. Our contributions to the field include:

- An overview of three prompting techniques we tested for classifying AI-related incidents by a predefined taxonomy of classes of harm, geographic location, societal impact, intended application, field of deployment, and others. We evaluate the three different techniques and the results and compare the LLM classifications to our human manual classifications.
- A classification of 17 incidents from AIID by LLM using various prompting techniques and another benchmark classification set by human manual efforts. Future efforts will aim to apply the most appropriate classification technique to the full 628 incidents on AIID and use those results for data visualization as part of the Responsible AI study.
- Evaluation and validation of the classification results by two domain experts, which assess the capability of the LLM in this classification task.

2 Related Work

2.1 AI Incidents Documentation

The documentation of AI failures has become increasingly crucial as AI technologies are more widely deployed in sensitive and critical contexts. To improve the quality of AI systems and prevent future incidents, the AI engineering community needs a comprehensive record of previous AI failures[10]. Various databases have been established to document AI incidents.

- **The AI and Algorithmic Incident and Controversies (AIAAIC) Repository** An independent initiative that logs global AI incidents and

controversies, managed by Charlie Pownall, and is used by over 60 universities and organizations [2].
- **The AI Vulnerability Database (AVID)** The database focuses on cataloging proven high-level failure modes and specific vulnerabilities, providing valuable insights for AI engineers and auditors [3].
- **Awful AI** A curated list to track current harmful usage of AI, aiming to raise awareness about the dark side of AI technologies [4].
- **AI Incidents Database (AIID)** The most utilized database for reporting AI incidents [5]. It encompasses a wide range of AI-related harms affecting physical health, safety, and social/political systems. This database also serves as the primary resource for our study.

Research using the records, including efforts to categorize the causes of AI failures [12] and a sector-based approach analysis of incidents [3], underscores the need for systematic approaches to classify and analyze these events. Despite the availability of databases cataloging numerous AI failures, similar incidents continue to occur. For instance, there are recurring instances where facial recognition systems disproportionately misidentify people of color, leading to wrongful accusations and arrests, despite widespread documentation and acknowledgment of bias in AI systems. Industry practitioners face significant challenges in effectively preventing and mitigating bias, often only identifying these serious issues after deployment [14].

These studies indicate that there is still much to learn and significant work to be done to develop a robust classification model capable of handling the complexity and variety of incident reports.

2.2 Large Language Models

With the advent of *pre-trained* language models (LM) with fixed architecture came a new paradigm shift in the learning of these models. LMs built on transformer architecture no longer rely on objective engineering nor fine-tuning to perform specific downstream tasks [9]. In the new paradigm, pre-trained LMs can solve various tasks with only textual *prompts* [9]. Previous works that have shown how LLMs can adapt and recognize tasks at inference time, also referred to as *in-context* learning [2]. These large models showed improved ability to learn a task from contextual information, where a model is guided by textual cues to perform a task. By processing an instruction or a handful of task examples, the model then predicts subsequent actions within the same context.

Specifically, in the domain of classification, LLMs have been applied with some success. For example, recent studies have deployed LLMs for classifying legal documents using detailed criteria [15] and have evaluated the GPT family of models for their effectiveness in biomedical reasoning and classification tasks

[2] https://www.aiaaic.org/aiaaic-repository
[3] https://avidml.org/
[4] https://github.com/daviddao/awful-ai
[5] https://github.com/responsible-ai-collaborative/aiid

[6]. The flexibility of LLMs to parse and organize complex data has facilitated their adoption in various burgeoning fields. However, relying entirely on LLM responses can pose risks, such as generating misleading answers in scenarios like academic forums, where models may produce incorrect responses if they fail to understand a question [18]. To mitigate these risks, type-specific prompting has been proposed to enhance the performance of GPT models in classification tasks. Our research builds upon this foundation by utilizing LLMs to automate the classification of AI incidents according to a predefined taxonomy, employing diverse prompting techniques to enhance this process.

3 Background

Previously introduced in Section 2.1, the dataset used in our study is sourced from the AI Incidents Database (AIID) GitHub repository. This dataset was chosen due to its extensive collection of incidents, featuring over 1,000 archived reports from more than 600 contributors across various sectors. These incidents document real-world cases where the use of AI has resulted in harm, covering a range of issues from facial recognition and targeted advertising to collisions involving autonomous vehicles, among many others.

Before introducing our task setup, we present the main classification fields that our LLM aims to identify. The predefined taxonomy for this study encompasses classes of AI-related harm ranging from discrimination, disinformation, human incompetence, pseudoscience, copyright violation, mental health implications, to environmental impact. In addition to the taxonomy of classes of harm, the LLM should also identify other fields of classification germane to our study of Responsible AI.

3.1 Classification Fields Explained

To clearly define the scope and detail of the data the LLM analyzes in each incident, we outline the specific fields it is designed to classify:

- **Geographical Location (Country, State, City, Continent):** These fields determine an incident's location, with the possibility of indicating "Worldwide" for events not confined to a single geographical area.
- **Company and Location (Company, Company City, Company State):** This identifies the organization responsible for the AI technology involved in the incident, including the precise location of its headquarters.
- **Affected Population:** This field aims to identify which demographic groups were impacted by the incident, shedding light on the societal segments most at risk from irresponsible AI usages.
- **Number of People Affected:** Both the actual and potential number of individuals impacted by the incident are quantified to gauge its severity and reach.

- **Taxonomic Classification of Classes of Harm (Classes, Subclasses, Sub-subclass):** Through a structured taxonomy, incidents are categorized into specific classes and subclasses of AI-related harm, enabling a systematic analysis of the types of irresponsible AI usage occurring. This taxonomy was collectively decided by the team and serve as labels or tags for the incident classification.
- **Application Area:** The specific domain of AI application involved in the incident is identified, such as healthcare, surveillance, or social media, providing insights into where AI poses the greatest risks.
- **Online:** Indicates whether the incident took place online, highlighting the digital nature of many AI-related ethical concerns.

Each classification field plays a crucial role in our comprehensive analysis of AI-related incidents from AIID, facilitating a nuanced understanding of how and where AI technologies are causing harm.

3.2 Taxonomy

We have adopted the comprehensive taxonomy developed by Dr. Ricardo Baeza-Yates as the classification framework for analyzing our dataset of incidents (see Figure 1). The taxonomy outlines the various classes of harm resulting from irresponsible AI use, forming a hierarchical structure where the main categories represent the broad types of harm, and the sub-classes offer a more detailed breakdown of these categories. For instance, 'Discrimination' serves as a main category that can be further dissected into 'Data Bias' or 'Algorithmic Bias.' Within 'Data Bias,' specifics such as gender, race, sexual orientation, economics, and others are identified as sub-subclass. This taxonomy will serve as the labels for the classes of harm in our classification.

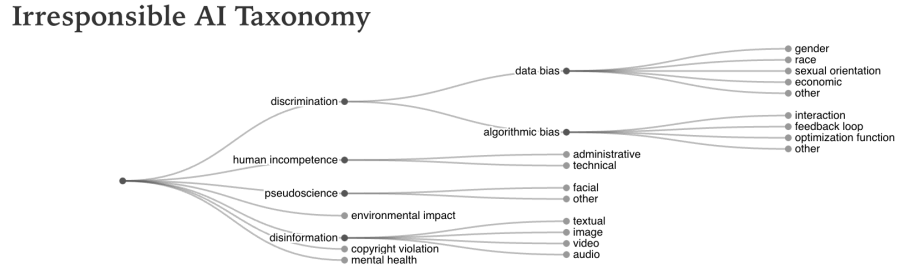

Fig. 1. Taxonomy for Classifying the Impacts of Irresponsible AI Utilization

Automation with Large Language Models (LLMs): We will program the LLMs to recognize and classify incidents according to the taxonomy. Their natural

language understanding will be leveraged to assess the incidents' attributes and categorize them accordingly.

Human Classification: Simultaneously, a manual classification process will be conducted. This will involve a thorough review of each incident by the members of the team who will then classify them based on the detailed criteria set out in the taxonomy. This process will not only validate the automated classifications but also ensure that nuanced or ambiguous cases are properly categorized.

4 Curation Methodologies

4.1 Experiment Setup

GPT-4 Model For our project, we utilized the Chat Completions API [6] by OpenAI to leverage the capabilities of an advanced language model for classifying AI-related incidents documented in our dataset. The model of choice for our classification task was the GPT-4 model [11]. The GPT-4 language model represents an advancement over its predecessor, GPT-3, featuring significantly more parameters. GPT-4's underlying technology is based on transformer architecture—specifically, attention mechanisms that emphasize different parts of the input data depending on the context. This model has been further refined with a technique called reinforcement learning from human feedback (RLHF), where human judgments help to fine-tune the model's responses [5]. Although we specifically selected the GPT-4 model for this study, our future work may involve working with different models and evaluating their performances comprehensively.

Input Data From our dataset, we concentrated on eliminating redundant entries from the dataset. We merged all URLs linked to a specific incident under a singular incident ID, as the same incident might be reported by multiple news sources. This de-duplication process significantly streamlined the dataset, ensuring that each unique incident was catalogued as a single, exhaustive entry.

The result is a curated dataset devoid of redundancies, comprising 628 incidents[7]. Each incident is distinctly identified by an incident ID, consolidating all pertinent URLs under this unique identifier. This enables the classification to be based on the substantive content of the articles associated with each incident.

Next, we aggregated all available URL links, drawing reports from various news sources, social media posts, or official statement, for each incident within the AIID database. Each URL serves as a crucial source of information for specific incidents, offering additional context, details, and diverse perspective. Using Newspaper3k article scraping library[8], we fetched the HTML content of each

[6] https://platform.openai.com/docs/api-reference/chat/create
[7] https://observablehq.com/@irresponsible-ai/aiid-data-processing
[8] https://newspaper.readthedocs.io/en/latest/

web-page, parsed it to extract relevant article text, and filtered out noise or irrelevant content. The extracted text was concatenated into a single comprehensive string for each incident, which then fed into the LLM as the input. By providing a substantial corpus of text data from multiple sources, the model is able to gain exposure to diverse language patterns, enhancing its ability to accurately understand and classify incidents.

4.2 GPT-4 Configurations

A critical parameter in our use of the Chat Completions API was the setting of the *temperature*, which influences the model's output variability. The temperature parameter can range from 0 to 1, with lower values producing more deterministic and less varied outputs, and higher values encouraging more creativity and diversity in the generated text [9]. For the scope of our study, which centered around the straightforward tasks of text interpretation and classification, we adjusted the temperature setting from 0.5 to 0.3. This adjustment was made with the intention of minimizing the model's propensity for generating creative responses, thereby ensuring that the output remained closely aligned with the factual content of the articles and the classification taxonomy.

4.3 Prompting Strategies

Delimiters The first few prompts did not have clear delimiters so sometimes the model was unable to understand the prompt instructions. Later we used delimiters to clearly indicate distinct parts of inputs[10]. This approach significantly enhanced the model's ability to comprehend and respond to the prompts as intended.

```
================== Start of Article Content ============
                    {article_text}
================== End of Article Content ==============
```

Listing 1.1. Using Delimiters in Prompts

Persona-Based Prompting To guide the behavior of the language model, we employed a strategy of persona adoption, where the model is instructed to assume a specific role within its operational framework. For instance, we asked the model to operate as "a helpful assistant designed to classify news articles into specific categories," with the expected output format being JSON. This approach is designed to orient the model towards a particular behavior pattern[11]

[9] https://platform.openai.com/docs/guides/text-generation/how-should-i-set-the-temperature-parameter

[10] https://platform.openai.com/docs/guides/prompt-engineering/tactic-use-delimiters-to-clearly-indicate-distinct-parts-of-the-input

[11] https://platform.openai.com/docs/guides/text-generation/chat-completions-api

Role Prompting for Enhanced Performance : The technique of role prompting leverages on the known sensitivity of Large Language Models (LLMs) to the construction of prompts. LLMs' performance can significantly benefit from the specification of roles within prompts, improving by at least 20 percent over control prompts where no specific context is provided [20]. This sensitivity underscores the effectiveness of role prompting in enhancing the model's task alignment and output accuracy.

We assigned an occupational role to our LLM. The initial prompt encouraged the LLM to conceptualize itself as part of a computer research institute tasked with categorizing incidents related to the irresponsible use of AI technology. Subsequently, we refined the prompting strategy by directly addressing our LLM as a computer researcher engaged in the same categorization task, eliminating the use of "imagine." This progression in prompt specificity was informed by evidence suggesting that LLMs respond with greater precision to direct role assignment, a reflection of their sensitivity to the nuances of prompt phrasing [20]. This iterative refinement in role prompting tailors the model's focus and improves its classification performance.

Detailed Step Instructions In the refined prompts, we detailed the steps necessary for task completion by specifying the actions required to achieve a desired outcome[12]. This modification improved GPT-4's compliance with the given instructions, a significant enhancement over earlier versions of the prompts that lacked such explicit direction. Previously, the model occasionally deviated from the prescribed format or introduced its own categories, despite explicit instructions to adhere to the predefined taxonomy. The incorporation of clearly defined steps and rules in the prompts resulted in outputs that more accurately reflected the intended instructions, demonstrating the LLM's increased ability to follow directions precisely. [13]

By adding step-by-step instructions refined the model's performance, the model was able to adhere to our instructions more closely and maintained a more consistent output format, whereas in the former zero-shot prompt some of the output format deviated from our specifications.

4.4 Prompting Techniques

Zero-Shot and Few-Shot Prompting The work of Kojima et al. [8] has shown that these LLMs possess remarkable capabilities for reasoning without prior exposure to specific data (zero-shot reasoning) and for rapidly adapting to tasks when provided with a small number of targeted examples (few-shot learning). The complexity of the task and capability of the model could influence how useful it is to include some examples in the prompt, as this may offer extra context and direction, potentially improving model performance [5]. However,

[12] https://platform.openai.com/docs/guides/prompt-engineering/tactic-specify-the-steps-required-to-complete-a-task

[13] See appendix A

Prompting without steps

Imagine a computer research institute trying to categorize a list of incidents of irresponsible use of artificial intelligence technology. Given the aggregated news article texts on relevant incidents, please extract the following information, your responses should be well thought-out and well-supported by the content of the articles, please also follow instructions of this prompt. Refer to {exampleOutput} for json output examples.

Prompting with added steps

1. Take a look at this taxonomy and take a close look at these examples
2. Your task is mainly this part. You have to fill out the following fields according to the article content using the three steps above:
STEP 1: Read the article text:
STEP 2: State your reason for your classifications for the following:
STEP 3: Make sure your classification output follows such format:

Fig. 2. Example of prompting without steps and prompting with added steps

it has also been shown that a well-designed zero-shot prompt, such as adding an intermediate reasoning step [8], could even lead to better performance and outperform few-shot prompts [5]. In our experiment, the few-shot learning with exemplars enabled GPT-4 to yield results that aligned more closely to the manual classifications[14].

Chain of Thought Chain-of-Thought (CoT) prompting is introduced in the work of Wei et al. [16] that by eliciting an intermediate natural language reasoning step, the prompt can expand the set of diverse tasks that LLMs can perform successfully. Other research literature have shown that LLMs, combined with in-context learning (ICL) and chain-of-thought (CoT) prompting, are capable of various high-level reasoning tasks. [8] also introduced by simply adding "Let's think step by step" to a zero-shot prompt, the model outperforms standard zero-shot prompting. In our classification task, we set out with standard zero-shot prompting and also zero-shot prompting combined with Chain-of-Thought to elicit reasoning steps. We find that in the case of zero-shot prompting, there was not a big difference in results whether or not we added the reasoning steps. However, adding reasoning steps in few-shot learning prompts significantly improved the classification outcomes.

Tree of Thoughts The Tree of Thoughts (ToT) concept was introduced in Yao et al. [17], which enables LLMs to perform deliberate problem-solving by

[14] See appendix B

maintaining and exploring a tree of thought steps. ToT frames problem-solving as a search over a tree of coherent language sequences ("thoughts") that serve as intermediate steps toward the solution, allowing exploration of multiple reasoning paths. This process bears resemblance to the thought and reasoning process typical to humans in that it encourages the LLM to plan and even backtrack in its reasoning to assess different paths of thoughts. It has been demonstrated that the ToT framework significantly outperforms standard input-output (IO) prompting and Chain-of-Thought prompting on certain reasoning and creative writing tasks [17].

Our classification task's prompting methodology draws inspiration from these research insights. We present our observations from employing various prompting strategies with modifications and ensemble approaches that combine select techniques.

5 Output

5.1 Zero-Shot Learning

In the zero-shot learning results, GPT-4 demonstrates proficiency in pinpointing geographic and location information within articles with considerable accuracy. However, it occasionally misidentifies the city associated with a company. For instance, it incorrectly recognized Stanford University's location as "Stanford" instead of the correct "Palo Alto" [15]. The "Affected population" category permitted a broader range of responses, given that it wasn't constrained to a predetermined set of labels. The model was successful in this field of classification by recognizing the impacted group of people. Additionally, GPT-4 was able to identify the hierarchy of harm classes along with their subcategories. Nevertheless, its categorization at times varied from those done manually [16].

5.2 Few-Shot Learning with Chain-of-Thought Technique

In our few-shot learning prompt,we included two examples in the prompt featuring article texts along with their example classification results. We also integrated the Chain of Thought (CoT) approach by providing a supporting rationale for each classification outcome. It has been shown that LLMs are capable of producing sequences of reasoning when provided with examples that showcase this process in the context of few-shot prompting [16]. The purpose of these examples is to guide the models in replicating a similar style of reasoning. By enumerating reasoning step examples, the model is guided to arrive at a particular answer by extracting information from the article. Eliciting reasoning steps from the model also allows it to correct its own answers [16]. Our findings reveal that GPT-4 was generally accurate in identifying geographic and location information, with the

[15] https://visit.stanford.edu/basics/

[16] See appendix E for the full result on the 17 incidents and appendix F for table comparisons

exception that the company city was sometimes misidentified. Figure 3 shows that location categories scored highest in accuracy when compared to manual classifications. GPT-4 also reliably classified "Affected population" and "Area of AI Application," which are categories that allow for a wider range of answers without fixed labels. With reasoning guidance provided in the examples, GPT-4's identification of the classes of harm were more consistent with manual classifications than with the results from zero-shot learning [17].

5.3 Tree of Thought Framework with Few-Shot Learning and Chain-of-Thought Technique

In addition to zero-shot and few-shot learning with CoT reasoning, we used an ensemble approach that combined few-shot learning and CoT reasoning with the Tree of Thoughts (ToT) framework. In the prompt, we asked GPT-4 to act as three experts in order to branch out into three reasoning and classification paths. The prompt guides the model to create a breadth-first search of the tree of probable classifications and ask the experts to vote on the classification is the most promising and well-supported by the article text and use that as the final result. Reference [17] mentions the "self-reflection" approach, wherein language models can provide feedback on the outputs they generate. The ToT framework leverages this capability to get the model to evaluate and select among its generated reasoning and outcome branches. We find GPT-4 performs generally accurate classification for geographic and location information, but again with exception of some misidentified company cities. The model also successfully identified the "Affected population" and "Area of AI Application" fields. The classifications for "Affected Population" provided by the ToT method included more detail compared to those from the Few-Shot Learning with Chain of Thought (CoT) approach. This variation could be attributed to the use of different sets of examples for the "Affected Population" category, with one set featuring concise, single-word examples and others offering descriptive examples of the affected demographic groups. It is uncertain which style or format the LLM will choose to emulate when provided with several different examples. When incorporating the Tree of Thoughts (ToT) into our prompts, the overall classification outcomes [18] were similar to those from few-shot learning with CoT, displaying only minor differences in certain areas.

6 Evaluation

Evaluation of the efficacy of prompt methods in LLMs are mainly divided to subjective and objective categories. Subjective evaluations generally rely on human evaluators to assess the quality of the generated content. Such evaluation methods are subjective and can be prone to inconsistencies [5]. In some cases, human

[17] See appendix C for full result on 17 incidents
[18] See appendix D

evaluation is the better way to evaluate the accuracy and quality of LLM generated results because evaluators (typically experts) can assess the results more comprehensively with more accurate feedback [4]. On the other hand, objective evaluation, or automatic evaluation methods include the use of machine learning algorithms to score the quality of the text generated by the LLMs. However, these automated metrics sometimes fail to fairly capture the full semantic considerations and cultural context in generated text as well as assessment results of human evaluators.

6.1 Objective Evaluation

We measured the accuracy score of the classification results from LLM from the manual results (used as the true labels). This evaluation metric focuses strictly on the literal exactness of the LLM's outputs, but it does not consider semantic meanings or contextual nuances. Consequently, only single-word fields such as such as "State", "Company State", "City", tend to score high with this metric. We find that the few-shot learning with ToT and CoT prompt results had the highest matches for these fields, as shown in figure 3. This suggests that few-shot learning with ToT and CoT has the closest reasoning process to our manual classifications. Other fields such as "Affected population", "Area of AI application", and the classes of harm are expected to have lower scores due to the literal nature of this accuracy measurement. In addition to accuracy, we used BERTScore [19], which measures semantic similarity between pairs of text sequences. It provides a more nuanced evaluation compared to taking the accuracy score or other traditional evaluation metrics that rely more on matching the exactness of n-grams. The F1 scores are relatively high across all fields for each prompting technique as seen in figure 4, suggesting that the BERT model is able to capture the similarity between the evaluated text and the reference. The few-shot with ToT and CoT results tends to have the highest or near-highest F1 score in almost every field compared to the other sources, which may suggest that this approach is better at aligning with the manual evaluation. Although BERTScore is better at capturing semantic nuances, it still does not take on full contextual and semantic meanings, which could account for the lower F1 scores across the classification fields for some categories.

6.2 Subjective Evaluation

Since our classification task is domain-specific to the study of Irresponsible AI and its impact, we used human subjective evaluation from domain experts to assess the quality of GPT-4's output. Building on the experimental framework outlined in Section 4, our evaluation method involved a blind review by two independent domain experts. Each expert assessed the classifications—both human and LLM-generated—across various dimensions such as geographic location, affected population, incident impact, application of AI use, and the classes of harm. The reviewers were unaware of which incidents have been classified by GPT-4 and which fields were manually classified by the team. The classifications

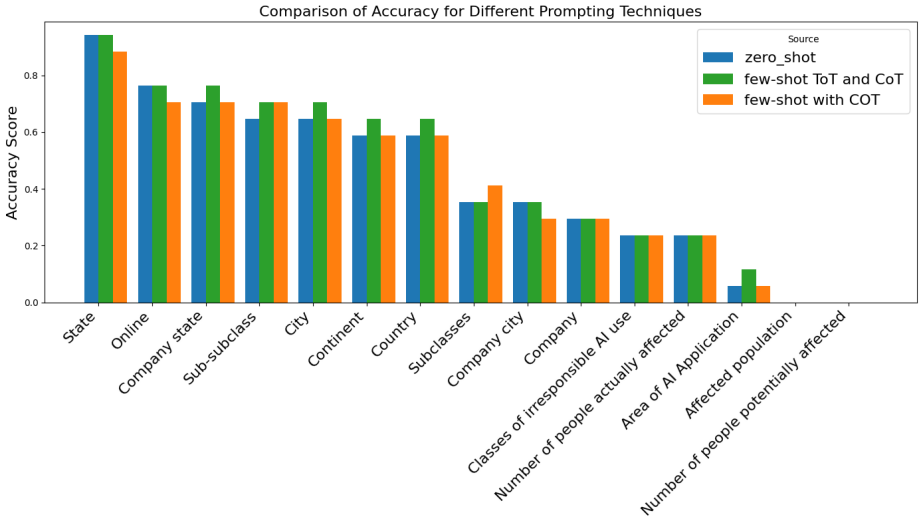

Fig. 3. Accuracy score of GPT-4 results. The accuracy scores here are measured by the exactness of literals using the manual classification results as the true label.

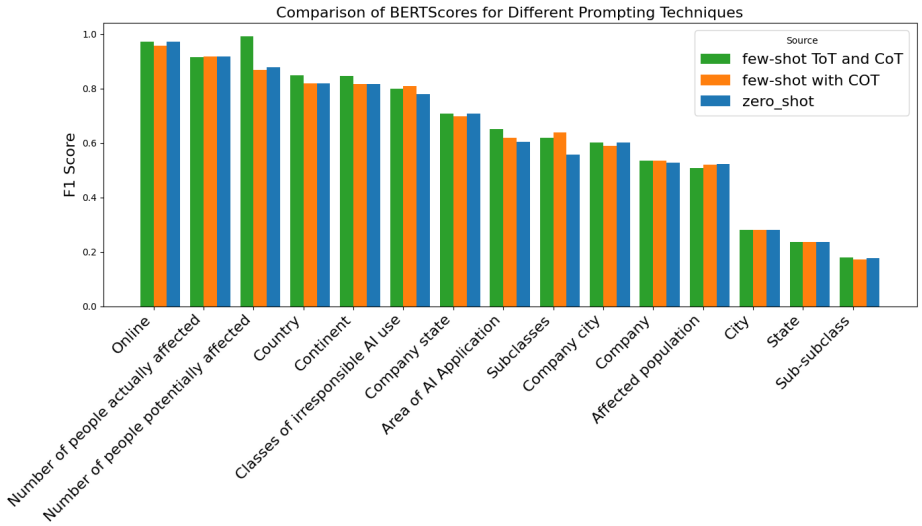

Fig. 4. F1 scores for different prompting technique results evaluated by the BERTScore evaluator

were scored on a scale from 1 to 10, based on accuracy of identification of each field. The average scores of the LLM classifications on 11 incidents achieved a slightly higher average score (at 7.73) than the average score of human manual classifications on 9 incidents (at 7.17). For both reviewers, the median scores for LLM classifications appear higher than for human classifications as can be seen in figures 5 and 6. This suggests that, on average, both reviewers rated LLM classifications more favorably than human ones. In both ratings, the variability of LLM classification ratings are generally wider than human classifications, indicating the reviewers both perceived some level of inconsistency in performance of the LLM results. Reviewer feedback on human classifications often cited issues such as typographical errors, incorrect identification of places, and lack of consensus regarding the labeling of harm classes, highlighting that human-generated results can also be prone to errors. On the other hand, the feedback for LLM classifications commonly mentioned the absence of predicted potential impacts, suggesting that the model may not have discerned certain information from the text or the information was not present for the model to make a judgment.

In some classification output, we have seen that GPT-4 is capable of producing very consistent output that is very close to human classification results [19]. Other cases show that the results varied with different prompts. The outcome of the assessment suggests that LLMs can outperform human efforts in classification tasks in some cases, as shown by the higher average scores. However they are not without limitations, such as failing to predict certain outcomes from information gleaned from article texts. The variability and errors in human classifications also demonstrate the complexities and challenges that we often see in the manual categorization processes. This comparison suggests that while LLMs show promise as efficient tools for classification, combining their strengths with human oversight could offer a more robust approach.

It is worth noting that this experiment represents a small subset of the 628 incidents documented in the AIID, meaning the findings might not sufficiently reflect the broader performance capabilities of LLM and human classifications. We plan to experiment in small batches before we proceed with a suitable approach that could successfully classify all the incidents. Additionally, the variability in evaluations among expert reviewers might indicate the need for a more robust evaluation design, since it is inevitable for subjective evaluations to have wide variability because of differences in opinions.

7 Conclusions

Our study's focus was to enhance classification methods for the numerous AI incidents documented in the AI Incidents Database, aiming to improve understanding and management of AI-related harm.

The study leveraged the advanced capabilities of large language models (LLMs) like GPT-4 to classify incidents more efficiently than manual methods. Our findings indicated that while LLMs can sometimes outperform human classification

[19] See appendix F

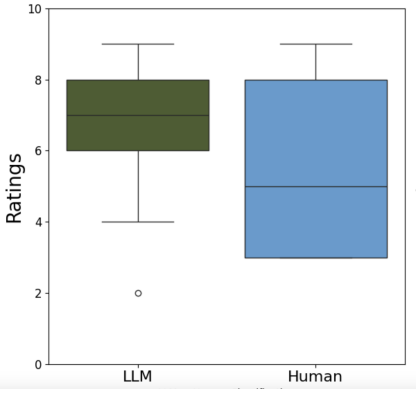

Fig. 5. Reviewer 1 ratings: distribution of reviewer 1's ratings on incidents classification from a blind review

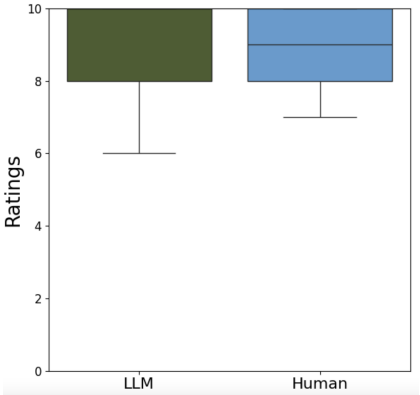

Fig. 6. Reviewer 2 ratings: distribution of reviewer 2's ratings on incidents classification from a blind review

in terms of average scoring, they are not infallible, occasionally misidentifying locations or missing potential impacts. This underscores the need for human oversight to ensure the highest quality and reliability in classification.

The investigation into prompt engineering, spanning zero-shot to few-shot learning with Chain of Thought and Tree of Thoughts frameworks, revealed nuances in model performance. Combining Tree of Thoughts method with few-shot prompts and explicit reasoning steps offered results most closely aligned with manual classifications. There are still known limitations to LLMs such as GPT-4. Most notably, like its predecessors, it still lacks complete reliability, sometimes producing fabricated information (known as "hallucinations") or committing logical errors [11]. Hallucinations occur because the model might not have enough supporting evidence in its training data for its responses, or it could generalize patterns too broadly to produce coherent output [7]. In some instances, we observed GPT-4 exhibiting this behavior where it could not accurately identify specific classification fields from the text of an article. For instance, it mistakenly identified the city of Stanford University as "Stanford" rather than "Palo Alto." In another case, it incorrectly classified "Silicon Valley" as the city where Chai Research [20] is headquartered, when it is actually located in Palo Alto [21].

7.1 Future Work

Moving forward, we plan to extend our methodology to the entire corpus of incidents in AIID and apply a more robust and effective approach to classify all the news incidents with high accuracy, which will be corroborated by experts in the field. Our aim is to harness these classifications to create data visualizations

[20] www.chai-research.com
[21] See appendix C

that will contribute to the ongoing discourse on Responsible AI, providing greater insight into AI-related harm trends and influencing the development of more responsible AI systems.

This work serves as a step toward understanding and documenting AI failures and establishing a robust, automated classification framework to handle the complexity of AI incident reports. It highlights both the capabilities and limitations of LLMs in this domain and points to a future where collaboration between human expertise and AI could offer robust solutions for managing AI's societal impacts.

Acknowledgments. We thank Khoury West Coast Research Program for funding this research and the team behind the AI Incidents Database (AIID) for publishing their data.

References

1. Baeza-Yates, R.: LECTURE HELD AT THE ACADEMIA EUROPAEA BUILDING BRIDGES CONFERENCE 2022: An Introduction to Responsible AI. European Review **31**(4), 406–421 (8 2023). https://doi.org/10.1017/S1062798723000145, https://www.cambridge.org/core/journals/european-review/article/lecture-held-at-the-academia-europaea-building-bridges-conference-2022/C61F82D2FF64FFB933282F5E65CDB22A
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., Mccandlish, S., Radford, A., Sutskever, I., Openai, D.A.: Language Models are Few-Shot Learners (2020)
3. Burema, D., Debowski-Weimann, N., Von Janowski, A., Grabowski, J., Maftei, M., Jacobs, M., Van Der Smagt, P., Benbouzid, D.: A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context. AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society pp. 705–714 (8 2023). https://doi.org/10.1145/3600211.3604680, https://dl.acm.org/doi/10.1145/3600211.3604680
4. Chang, Y., Wang, X.U., Yi, X., Wang, Y., Ye, W., Yu, P.S., Chang, Y., Wang, X., Wu, Y., Yi, X., Xie, X., Yang, .L., Wang, C., Zhang, Y., Wang, Y., Ye, W., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Wang, C., Zhang, Y., Chang, Y., Yang, Q., Xie, X.: A Survey on Evaluation of Large Language Models. ACM Trans. Intell. Syst. Technol **15**(3), 39 (2024). https://doi.org/10.1145/3641289, https://github.com/MLGroupJLU/LLM-eval-survey
5. Chen, B., Zhang, Z., Langrené, N., Zhu, S.: Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review (10 2023), http://arxiv.org/abs/2310.14735

6. Chen, S., Li, Y., Lu, S., Van, H., Aerts, H.J.W.L., Savova, G.K., Bitterman, D.S.: Evaluating the ChatGPT family of models for biomedical reasoning and classification. Journal of the American Medical Informatics Association (4 2024). https://doi.org/10.1093/JAMIA/OCAD256
7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of Hallucination in Natural Language Generation. ACM Computing Surveys **55**(12) (3 2023). https://doi.org/10.1145/3571730, https://dl.acm.org/doi/10.1145/3571730
8. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large Language Models are Zero-Shot Reasoners (5 2022), http://arxiv.org/abs/2205.11916
9. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (2021)
10. Mcgregor, S.: Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database (2021), www.aaai.org
11. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, , Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, , Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F.d.A.B., Petrov, M., Pinto, H.P.d.O., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet,

P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: GPT-4 Technical Report (3 2023), http://arxiv.org/abs/2303.08774

12. Pittaras, N., McGregor, S.: A taxonomic system for failure cause analysis of open source AI incidents (11 2022), https://arxiv.org/abs/2211.07280v1
13. Saparov, A., Pang, R.Y., Padmakumar, V., Joshi, N., Kazemi, S.M., Kim, N., He, H.: Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples (5 2023), http://arxiv.org/abs/2305.15269
14. Turri, V., Dzombak, R.: Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society pp. 576–583 (8 2023). https://doi.org/10.1145/3600211.3604700, https://dl.acm.org/doi/10.1145/3600211.3604700
15. Wei, F., Keeling, R., Huber-Fliflet, N., Zhang, J., Dabrowski, A., Yang, J., Mao, Q., Qin, H.: Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review. Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023 pp. 2786–2792 (2023). https://doi.org/10.1109/BIGDATA59044.2023.10386911
16. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (1 2022), http://arxiv.org/abs/2201.11903
17. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of Thoughts: Deliberate Problem Solving with Large Language Models (5 2023), http://arxiv.org/abs/2305.10601
18. Zhang, P., Jaipersaud, B., Ba, J., Petersen, A., Zhang, L., Zhang, M.R.: Classifying Course Discussion Board Questions using LLMs. Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE **2**, 658 (6 2023). https://doi.org/10.1145/3587103.3594202
19. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT (4 2019), http://arxiv.org/abs/1904.09675
20. Zheng, M., Pei, J., Jurgens, D.: Is "A Helpful Assistant" the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts. arxiv.orgM Zheng, J Pei, D JurgensarXiv preprint arXiv:2311.10054, 2023•arxiv.org (11 2023), http://arxiv.org/abs/2311.10054

2

Appendices

A Prompt with clear steps

```
STEP 1: Read the article text:
================== Start of Article Content =================
{article_text}
================== End of Article Content ====================
STEP 2: State your reason for your classifications for the
    following:
=================Classification Fields====================
- Country (output "Worldwide" if the incident happened across
     multiple countries):
- State (if not applicable leave blank):
- City (if not applicable leave blank):
- Continent (output "Worldwide" if the incident happened
    across multiple countries):
- Company (i.e. the company that developed the technology
    involved in this incident):
- Company city (the city where the headquarters of this
    company is located. If the company recently moved
    headquarters, please use the location of the new
    headquarter):
- Company state (the state of the company city, if applicable
    , if not leave blank):
- Affected population (think about which groups of people are
     directly affected by the incident in the article.)
- Number of people actually affected (let's check the number
    of people directly affected according to the article.
    Give a total number. If unknown output 'Unknown'):
- Number of people potentially affected (let's check the
    article text to see if this information is provided or
    suggested, if not you may ouput 'Unknown'):
- Classes of irresponsible AI use (please follow the rules
    and refer to this taxonomy:
'''taxonomy classes
        {taxa.classes}
'''
Rule1: There could be more than one classes the article
    classifies as.
Rule2: DO NOT create your own class, adhere strictly to the
    provided list.
```

```
- Subclasses (please follow the rules and refer to this
    taxonomy structure ‘<class>:[<subclass>]‘):
  ‘‘‘taxonomy subclasses
        {taxa.subclasses}
  ‘‘‘
==================Classification Fields====================
STEP 3: Make sure your classification output follows such
    format:
‘‘‘THIS IS AN EXAMPLE‘‘‘
{example_output_id_6}
‘‘‘END OF EXAMPLE‘‘‘
```

Listing 2.1. Detailed Instructions with Steps in Prompts

B Manual classification results

The following JSON of 17 incidents are the results of our manual classifications that was discussed and decided among three people after reading the articles of the incidents. This was used as a benchmark for comparison with the LLM generated results.


```
[
  {
    "id": 1,
    "Country": "Worldwide",
    "State": "",
    "City": "",
    "Continent": "Worldwide",
    "Company": "Google LLC",
    "Company city": "Mountain View",
    "Company state": "California",
    "Affected population": "Children on Youtube",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "human incompetence,
    mental health,copyright",
    "Subclasses": "technical,administrative",
    "Sub-subclass": "",
    "Area of AI Application": "content filtering",
    "Online": "yes"
  },
  {
    "id": 2,
    "Country": "United States",
    "State": "New Jersey",
    "City": "Robbinsville",
    "Continent": "North America",
```

```
    "Company": "Amazon",
    "Company city": "Seattle",
    "Company state": "Washington",
    "Affected population": "Amazon Workers and Families",
    "Number of people actually affected": 54,
    "Number of people potentially affected": 80,
    "Classes of irresponsible AI use": "human incompetence",
    "Subclasses": "technical",
    "Sub-subclass": "",
    "Area of AI Application": "robotics",
    "Online": "no"
  },
  {
    "id": 3,
    "Country": "Indonesia",
    "State": "",
    "City": "Bali",
    "Continent": "Southeast Asia",
    "Company": "Boeing",
    "Company city": "Arlington",
    "Company state": "Virginia",
    "Affected population": "Lion Air Jet Passengers and
   Families",
    "Number of people actually affected": 189,
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "human incompetence",
    "Subclasses": "technical",
    "Sub-subclass": "",
    "Area of AI Application": "airspeed indicator",
    "Online": "no"
  },
  {
    "id": 4,
    "Country": "United States",
    "State": "Arizona",
    "City": "Tempe",
    "Continent": "North America",
    "Company": "Uber",
    "Company city": "San Francisco",
    "Company state": "California",
    "Affected population": "Pedestrians",
    "Number of people actually affected": 1,
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "human incompetence",
    "Subclasses": "technical",
    "Sub-subclass": "",
    "Area of AI Application": "autonomous driving",
    "Online": "no"
  },
  {
```

```
    "id": 5,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "",
    "Company city": "",
    "Company state": "",
    "Affected population": "Patients with robotic procedures
    ",
    "Number of people actually affected": 1535,
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "human incompetence",
    "Subclasses": "technical",
    "Sub-subclass": "",
    "Area of AI Application": "robotics ",
    "Online": "no"
  },
  {
    "id": 6,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "Microsoft",
    "Company city": "Seattle",
    "Company state": "Washington",
    "Affected population": "Tweeter users",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "discrimination,
    disinformation",
    "Subclasses": "data bias,algorithmic bias,textual",
    "Sub-subclass": "race,gender,feedback loop",
    "Area of AI Application": "chatbot",
    "Online": "yes"
  },
  {
    "id": 9,
    "Country": "United States",
    "State": "New York",
    "City": "New York",
    "Continent": "North America",
    "Company": "",
    "Company city": "New York",
    "Company state": "New York",
    "Affected population": "NYC teachers",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
```

```
    "Classes of irresponsible AI use": "discrimination,
   disinformation",
    "Subclasses": "data bias,algorithmic bias,textual",
    "Sub-subclass": "other,other",
    "Area of AI Application": "data prediction",
    "Online": "no"
  },
  {
    "id": 10,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "UKG",
    "Company city": "Lowell",
    "Company state": "Massachusetts",
    "Affected population": "Starbucks braristas",
    "Number of people actually affected": "130,000",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "human incompetence",
    "Subclasses": "technical",
    "Sub-subclass": "",
    "Area of AI Application": "job scheduling",
    "Online": "no"
  },
  {
    "id": 11,
    "Country": "United States",
    "State": "Florida",
    "City": "Broward County",
    "Continent": "North America",
    "Company": "Northpointe Bank",
    "Company city": "Grand Rapids",
    "Company state": "Michigan",
    "Affected population": "Defendants in Broward County,",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "discrimination,
   pseudoscience",
    "Subclasses": "data bias,facial",
    "Sub-subclass": "race",
    "Area of AI Application": "predictive algorithms - COMPAS
    ",
    "Online": "no"
  },
  {
    "id": 13,
    "Country": "United States",
    "State": "",
    "City": "",
```

```
    "Continent": "North America",
    "Company": "Google LLC",
    "Company city": "Mountain View",
    "Company state": "California",
    "Affected population": "Miniorties and groups like woman,
     gay, etc.",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "discrimination",
    "Subclasses": "data bias",
    "Sub-subclass": "gender,race,sexual orientation,other",
    "Area of AI Application": "predictive algorithm",
    "Online": "yes"
  },
  {
    "id": 14,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "Google LLC",
    "Company city": "Mountain View",
    "Company state": "California",
    "Affected population": "Miniorties and groups like woman,
     gay, etc.",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "discrimination",
    "Subclasses": "algorithmic bias,data bias",
    "Sub-subclass": "feedback loop,race,gender",
    "Area of AI Application": "NLP analysis",
    "Online": "yes"
  },
  {
    "id": 451,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "Stability AI",
    "Company city": "Houston",
    "Company state": "Texas",
    "Affected population": "Getty Company",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "copyright violation",
    "Subclasses": "",
    "Sub-subclass": "",
    "Area of AI Application": "image generation",
    "Online": "yes"
```

```
  },
  {
    "id": 382,
    "Country": "United Kingdom",
    "State": "",
    "City": "London",
    "Continent": "Europe",
    "Company": "Facebook",
    "Company city": "Menlo Park",
    "Company state": "California",
    "Affected population": "instagram users",
    "Number of people actually affected": 1,
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "mental health",
    "Subclasses": "",
    "Sub-subclass": "",
    "Area of AI Application": "social media",
    "Online": "yes"
  },
  {
    "id": 505,
    "Country": "Belgium",
    "State": "",
    "City": "",
    "Continent": "Europe",
    "Company": "EleutherAI",
    "Company city": "New York",
    "Company state": "New York",
    "Affected population": "Chai app users",
    "Number of people actually affected": 1,
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "mental health",
    "Subclasses": "",
    "Sub-subclass": "",
    "Area of AI Application": "chatbot",
    "Online": "yes"
  },
  {
    "id": 167,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "Standford researchers",
    "Company city": "Palo Alto",
    "Company state": "California",
    "Affected population": "LGBT group",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
```

```
    "Classes of irresponsible AI use": "pseudoscience,
   discrimination",
    "Subclasses": "facial,data bias",
    "Sub-subclass": "sexual orientation",
    "Area of AI Application": "behavioral modeling",
    "Online": "yes"
  },
  {
    "id": 0,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "",
    "Company city": "",
    "Company state": "",
    "Affected population": "General Human",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "environmental impact
   ",
    "Subclasses": "",
    "Sub-subclass": "",
    "Area of AI Application": "AI training",
    "Online": "no"
  },
  {
    "id": 39,
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "",
    "Company city": "",
    "Company state": "",
    "Affected population": "online audience",
    "Number of people actually affected": "unknown",
    "Number of people potentially affected": "unknown",
    "Classes of irresponsible AI use": "disinformation",
    "Subclasses": "video",
    "Sub-subclass": "",
    "Area of AI Application": "deepfake video generation",
    "Online": "yes"
  }
]
```


Listing 2.2. Manual classification results of 17 incidents handpicked by the team. The 17 incidents covered all the classes of harm of our taxonomy.

C Few-Shot Learning Classification Results

These are the classification results of 17 incidents from GPT-4 with two-shot prompting with Chain-of-Thought (CoT) technique.


```
{
    "1": {
        "Country": "Global",
        "State": "N/A",
        "City": "N/A",
        "Continent": "Global",
        "Company": "Google LLC",
        "Company city": "Mountain View, CA",
        "Affected population": "Children",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Class of irresponsible AI use": [
            "Disinformation",
            "Human Incompetence",
            "Mental Health",
            "Other"
        ],
        "Subclasses": [
            "Disinformation-> Video",
            "Human Incompetence-> Technical",
            "Mental Health-> Psychological Impact",
            "Other-> Inappropriate Content"
        ],
        "Sub-subclass": [],
        "Area of AI Application": "Video Sharing/Streaming",
        "Online": "Yes"
    },
    "2": {
        "Country": "United States",
        "State": "New Jersey",
        "City": "Robbinsville",
        "Continent": "North America",
        "Company": "Amazon",
        "Company city": "Seattle, WA",
        "Affected population": "Amazon workers",
        "Number of people actually affected": "24",
        "Number of people potentially affected": "54",
        "Class of irresponsible AI use": [
            "Human Incompetence"
        ],
        "Subclasses": [
            "Human Incompetence-> Technical"
        ],
        "Sub-subclass": [],
```

```
        "Area of AI Application": "warehouse automation, robotics",
        "Online": "No"
    },
    "3": {
        "Country": "Indonesia",
        "State": "",
        "City": "Jakarta",
        "Continent": "Asia",
        "Company": "Boeing",
        "Company city": "Chicago, IL",
        "Affected population": "Airline passengers and crew",
        "Number of people actually affected": "189",
        "Number of people potentially affected": "Unknown",
        "Class of irresponsible AI use": [
            "Human Incompetence"
        ],
        "Subclasses": [
            "Human Incompetence-> Technical"
        ],
        "Sub-subclass": [],
        "Area of AI Application": "Aviation safety systems",
        "Online": "No"
    },
    "4": {
        "Country": "United States",
        "State": "Arizona",
        "City": "Tempe",
        "Continent": "North America",
        "Company": "Uber",
        "Company city": "San Francisco, CA",
        "Affected population": "Pedestrians",
        "Number of people actually affected": "1",
        "Number of people potentially affected": "Unknown",
        "Class of irresponsible AI use": [
            "Human Incompetence",
            "Technical Failure"
        ],
        "Subclasses": [
            "Human Incompetence-> Monitoring Failure",
            "Technical Failure-> Detection Failure"
        ],
        "Sub-subclass": [],
        "Area of AI Application": "autonomous vehicles",
        "Online": "No"
    },
    "5": {
        "Country": "United States",
        "State": "Not specified",
        "City": "Not specified",
```

```
        "Continent": "North America",
        "Company": "Not specified",
        "Company city": "Not specified",
        "Affected population": "Patients undergoing robotic surgery",
        "Number of people actually affected": "144 deaths, 1391 injuries",
        "Number of people potentially affected": "More than 1.7 million (number of robotic procedures carried out between 2007 and 2013)",
        "Class of irresponsible AI use": [
            "Human Incompetence",
            "Technical Failure"
        ],
        "Subclasses": [
            "Human Incompetence-> Operator error",
            "Technical Failure-> Device malfunction",
            "Technical Failure-> System error"
        ],
        "Sub-subclass": [
            "Device malfunction-> Electrical sparking",
            "Device malfunction-> Uncontrolled movement",
            "Device malfunction-> Broken pieces",
            "System error-> Loss of video feed"
        ],
        "Area of AI Application": "robotic surgery",
        "Online": "No"
    },
    "6": {
        "Country": "Global",
        "State": "Not specified",
        "City": "Not specified",
        "Continent": "Not specified",
        "Company": "Microsoft",
        "Company city": "Redmond, WA",
        "Affected population": "Twitter Users, Online Community",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Millions (Twitter user base and wider online community)",
        "Class of irresponsible AI use": [
            "Human Incompetence",
            "Disinformation"
        ],
        "Subclasses": [
            "Human Incompetence-> Lack of foresight",
            "Disinformation-> Racial",
            "Disinformation-> Political",
            "Disinformation-> Gender"
        ],
```

```
        "Sub-subclass": [],
        "Area of AI Application": "Social media interaction, Conversational AI",
        "Online": "Yes"
    },
    "7": {
        "Country": "Global",
        "State": "Global",
        "City": "Global",
        "Continent": "Global",
        "Company": "Wikipedia",
        "Company city": "San Francisco, CA",
        "Affected population": "Wikipedia Users",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Millions",
        "Class of irresponsible AI use": [
            "Human Incompetence"
        ],
        "Subclasses": [
            "Human Incompetence-> Technical"
        ],
        "Sub-subclass": [],
        "Area of AI Application": "Online encyclopedia editing",
        "Online": "Yes"
    },
    "8": {
        "Country": "United States",
        "State": "California",
        "City": "San Francisco",
        "Continent": "North America",
        "Company": "Uber Technologies Inc.",
        "Company city": "San Francisco, CA",
        "Affected population": "General Public",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Residents and visitors of San Francisco",
        "Class of irresponsible AI use": [
            "Human Incompetence",
            "Other"
        ],
        "Subclasses": [
            "Human Incompetence-> Technical",
            "Other-> Safety"
        ],
        "Sub-subclass": [],
        "Area of AI Application": "autonomous vehicles",
        "Online": "No"
    },
    "9": {
```

```
        "Country": "United States",
        "State": "New York",
        "City": "New York City",
        "Continent": "North America",
        "Company": "New York City Department of Education",
        "Company city": "New York, NY",
        "Affected population": "Teachers",
        "Number of people actually affected": "12000",
        "Number of people potentially affected": "Unknown",
        "Class of irresponsible AI use": [
            "Human Incompetence",
            "Other"
        ],
        "Subclasses": [
            "Human Incompetence-> Technical",
            "Other-> Misuse of Data"
        ],
        "Sub-subclass": [],
        "Area of AI Application": "Educational assessment",
        "Online": "Yes"
    },
    "10": {
        "Country": "United States",
        "State": "Various",
        "City": "Various",
        "Continent": "North America",
        "Company": "Starbucks",
        "Company city": "Seattle, WA",
        "Affected population": "Starbucks Workers",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "130,000 (
    total number of Starbucks baristas in the U.S.)",
        "Class of irresponsible AI use": [
            "Human Incompetence"
        ],
        "Subclasses": [
            "Human Incompetence-> Technical"
        ],
        "Sub-subclass": [],
        "Area of AI Application": "workforce management,
    scheduling",
        "Online": "No"
    },
    "11": {
        "Country": "United States",
        "State": "Florida",
        "City": "Broward County",
        "Continent": "North America",
        "Company": "Northpointe (now part of Equivant)",
        "Company city": "Canton, OH",
```

```
        "Affected population": "Criminal Defendants",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Class of irresponsible AI use": [
            "Discrimination",
            "Human Incompetence"
        ],
        "Subclasses": [
            "Discrimination-> Data bias",
            "Discrimination-> Algorithmic bias",
            "Human Incompetence-> Technical"
        ],
        "Sub-subclass": [
            "Data bias-> Race",
            "Data bias-> Gender",
            "Algorithmic bias-> Race",
            "Algorithmic bias-> Gender"
        ],
        "Area of AI Application": "criminal risk assessment",
        "Online": "Yes"
    },
    "12": {
        "Country": "",
        "State": "",
        "City": "",
        "Continent": "",
        "Company": "",
        "Company city": "",
        "Affected population": "",
        "Number of people actually affected": "",
        "Number of people potentially affected": "",
        "Class of irresponsible AI use": [],
        "Subclasses": [],
        "Sub-subclass": [],
        "Area of AI Application": "",
        "Online": ""
    },
    "13": {
        "Country": "Global",
        "State": "N/A",
        "City": "N/A",
        "Continent": "Global",
        "Company": "Google LLC",
        "Company city": "Mountain View, CA",
        "Affected population": "Online Users",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Millions",
        "Class of irresponsible AI use": [
            "Discrimination"
        ],
```

```
        "Subclasses": [
            "Discrimination-> Data bias",
            "Discrimination-> Algorithmic bias"
        ],
        "Sub-subclass": [
            "Data bias-> Race",
            "Data bias-> Sexual Orientation",
            "Data bias-> Economic",
            "Algorithmic bias-> Gender",
            "Algorithmic bias-> Race"
        ],
        "Area of AI Application": "content moderation, online
     discussion forums",
        "Online": "Yes"
    },
    "14": {
        "Country": "Global",
        "State": "Global",
        "City": "Global",
        "Continent": "Global",
        "Company": "Google LLC",
        "Company city": "Mountain View, CA",
        "Affected population": "Users of Google's Cloud
    Natural Language API, including developers and indirectly
    , religious and ethnic minorities",
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Class of irresponsible AI use": [
            "Discrimination"
        ],
        "Subclasses": [
            "Discrimination-> Data bias",
            "Discrimination-> Algorithmic bias"
        ],
        "Sub-subclass": [
            "Data bias-> Race",
            "Data bias-> Sexual Orientation"
        ],
        "Area of AI Application": "Natural Language
    Processing, Sentiment Analysis",
        "Online": "Yes"
    },
    "15": {
        "1": {
            "Country": "Global",
            "State": "",
            "City": "",
            "Continent": "",
            "Company": "Amazon",
            "Company city": "Seattle, WA",
```

```
            "Affected population": "Authors and readers of
    gay and lesbian-themed books",
            "Number of people actually affected": "Unknown",
            "Number of people potentially affected": "Unknown
    ",
            "Class of irresponsible AI use": [
                "Human Incompetence"
            ],
            "Subclasses": [
                "Human Incompetence-> Technical"
            ],
            "Sub-subclass": [],
            "Area of AI Application": "Online retail, Sales
    ranking system",
            "Online": "Yes"
        }
    }
}
```

Listing 2.3. Few-shot with CoT classification results of 17 incidents

D Tree of Thought Framework Classification Results

The Tree of Thoughts framework allows the LLM to explore various reasoning paths and collectively determine the most well-supported classification based on the text of the article text.

```
{
    "1": {
        "Country": "Worldwide",
        "State": "",
        "City": "",
        "Continent": "Worldwide",
        "Company": "YouTube",
        "Company city": "San Bruno",
        "Company state": "California",
        "Affected population": [
            "Student",
            "Online Female Population",
            "Twitter Users"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Disinformation",
            "Mental Health",
            "Other"
```

```
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ],
            "Disinformation": [
                "Textual",
                "Image",
                "Video"
            ],
            "Mental Health": [],
            "Other": []
        },
        "Sub-subclass": [],
        "Area of AI Application": "content filtering",
        "Online": "yes"
    },
    "2": {
        "Country": "United States",
        "State": "New Jersey",
        "City": "Robbinsville",
        "Continent": "North America",
        "Company": "Amazon",
        "Company city": "Seattle",
        "Company state": "Washington",
        "Affected population": [
            "Warehouse Workers"
        ],
        "Number of people actually affected": "24",
        "Number of people potentially affected": "54",
        "Classes of irresponsible AI use": [
            "Human Incompetence"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ]
        },
        "Sub-subclass": [],
        "Area of AI Application": "Warehouse automation",
        "Online": "no"
    },
    "3": {
        "Country": "Indonesia",
        "State": "",
        "City": "Jakarta",
        "Continent": "Asia",
        "Company": "Boeing",
        "Company city": "Chicago",
        "Company state": "Illinois",
```

```
        "Affected population": [
            "Air Travelers"
        ],
        "Number of people actually affected": "189",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Other"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ],
            "Other": []
        },
        "Sub-subclass": {
            "Technical": []
        },
        "Area of AI Application": "Aviation safety and
    control systems",
        "Online": "no"
    },
    "4": {
        "Country": "USA",
        "State": "Arizona",
        "City": "Tempe",
        "Continent": "North America",
        "Company": "Uber",
        "Company city": "San Francisco",
        "Company state": "California",
        "Affected population": [
            "Pedestrians"
        ],
        "Number of people actually affected": "1",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Other"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ],
            "Other": []
        },
        "Sub-subclass": [],
        "Area of AI Application": "Autonomous Vehicles",
        "Online": "no"
    },
    "5": {
```

```
        "Country": "United States",
        "State": "",
        "City": "",
        "Continent": "North America",
        "Company": "Not specified",
        "Company city": "Not specified",
        "Company state": "",
        "Affected population": [
            "Patients undergoing robotic surgery"
        ],
        "Number of people actually affected": "1535",
        "Number of people potentially affected": "1.7 million
    ",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Technical"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ]
        },
        "Sub-subclass": [],
        "Area of AI Application": "Robotic surgery",
        "Online": "no"
    },
    "6": {
        "Country": "Worldwide",
        "State": "",
        "City": "",
        "Continent": "Worldwide",
        "Company": "Microsoft",
        "Company city": "Redmond",
        "Company state": "Washington",
        "Affected population": [
            "Twitter Users",
            "Online Population"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Disinformation",
            "Other"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ],
            "Disinformation": [
```

```
                "Textual"
            ],
            "Other": []
        },
        "Sub-subclass": [],
        "Area of AI Application": "Social Media Interaction",
        "Online": "yes"
    },
    "7": {
        "Country": "Worldwide",
        "State": "",
        "City": "",
        "Continent": "Worldwide",
        "Company": "Wikipedia",
        "Company city": "San Francisco",
        "Company state": "California",
        "Affected population": [
            "Online Users"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Other"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ],
            "Other": []
        },
        "Sub-subclass": [],
        "Area of AI Application": "Content management and
    editing",
        "Online": "yes"
    },
    "8": {
        "Country": "United States",
        "State": "California",
        "City": "San Francisco",
        "Continent": "North America",
        "Company": "Uber",
        "Company city": "San Francisco",
        "Company state": "California",
        "Affected population": [
            "General Public",
            "Bicyclists",
            "Pedestrians"
        ],
        "Number of people actually affected": "Unknown",
```

```
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Other"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ],
            "Other": []
        },
        "Sub-subclass": {
            "Technical": []
        },
        "Area of AI Application": "Autonomous Vehicles",
        "Online": "no"
    },
    "9": {
        "Country": "United States",
        "State": "New York",
        "City": "New York City",
        "Continent": "North America",
        "Company": "New York City Department of Education",
        "Company city": "New York City",
        "Company state": "New York",
        "Affected population": [
            "Student",
            "Teachers"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Other"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Administrative"
            ],
            "Other": []
        },
        "Sub-subclass": [],
        "Area of AI Application": "Educational assessment",
        "Online": "yes"
    },
    "10": {
        "Country": "United States",
        "State": "",
        "City": "",
        "Continent": "North America",
```

```
        "Company": "Starbucks",
        "Company city": "Seattle",
        "Company state": "Washington",
        "Affected population": [
            "Retail Workers",
            "Single Parents",
            "Students"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Human Incompetence",
            "Discrimination"
        ],
        "Subclasses": {
            "Human Incompetence": [
                "Technical"
            ],
            "Discrimination": [
                "Data bias"
            ]
        },
        "Sub-subclass": {
            "Technical": [],
            "Data bias": [
                "Economic",
                "Other"
            ]
        },
        "Area of AI Application": "Workforce management",
        "Online": "yes"
    },
    "11": {
        "Country": "United States",
        "State": "Florida",
        "City": "Broward County",
        "Continent": "North America",
        "Company": "Northpointe (now part of Equivant)",
        "Company city": "Canton",
        "Company state": "Ohio",
        "Affected population": [
            "Black Defendants",
            "White Defendants"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Discrimination"
        ],
        "Subclasses": {
```

```
            "Discrimination": [
                "Data bias",
                "Algorithmic bias"
            ]
        },
        "Sub-subclass": {
            "Data bias": [
                "Race",
                "Gender"
            ],
            "Algorithmic bias": [
                "Feedback loop",
                "Optimization function"
            ]
        },
        "Area of AI Application": "Criminal justice risk
    assessments",
        "Online": "yes"
    },
    "12": {
        "Country": "",
        "State": "",
        "City": "",
        "Continent": "",
        "Company": "",
        "Company city": "",
        "Company state": "",
        "Affected population": [],
        "Number of people actually affected": "",
        "Number of people potentially affected": "",
        "Classes of irresponsible AI use": [],
        "Subclasses": {},
        "Sub-subclass": [],
        "Area of AI Application": "",
        "Online": ""
    },
    "13": {
        "Country": "United States",
        "State": "",
        "City": "",
        "Continent": "North America",
        "Company": "Google's Alphabet",
        "Company city": "Mountain View",
        "Company state": "California",
        "Affected population": [
            "Online Population"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
```

```
            "Discrimination",
            "Human Incompetence"
        ],
        "Subclasses": {
            "Discrimination": [
                "Data bias",
                "Algorithmic bias"
            ],
            "Human Incompetence": [
                "Technical"
            ]
        },
        "Sub-subclass": {
            "Data bias": [
                "Gender",
                "Race",
                "Sexual Orientation"
            ],
            "Algorithmic bias": [
                "Other"
            ]
        },
        "Area of AI Application": "content filtering",
        "Online": "yes"
    },
    "14": {
        "Country": "Worldwide",
        "State": "",
        "City": "",
        "Continent": "Worldwide",
        "Company": "Google",
        "Company city": "Mountain View",
        "Company state": "California",
        "Affected population": [
            "Online Population"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Discrimination"
        ],
        "Subclasses": {
            "Discrimination": [
                "Data bias"
            ]
        },
        "Sub-subclass": {
            "Data bias": [
                "Gender",
                "Race",
```

```
                "Sexual Orientation"
            ]
        },
        "Area of AI Application": "Natural Language
    Processing",
        "Online": "yes"
    },
    "15": {
        "Country": "Worldwide",
        "State": "",
        "City": "",
        "Continent": "Worldwide",
        "Company": "Amazon.com",
        "Company city": "Seattle",
        "Company state": "Washington",
        "Affected population": [
            "LGBTQ"
        ],
        "Number of people actually affected": "Unknown",
        "Number of people potentially affected": "Unknown",
        "Classes of irresponsible AI use": [
            "Discrimination",
            "Human Incompetence"
        ],
        "Subclasses": {
            "Discrimination": [
                "Data bias"
            ],
            "Human Incompetence": [
                "Technical"
            ]
        },
        "Sub-subclass": {
            "Data bias": [
                "Sexual Orientation"
            ]
        },
        "Area of AI Application": "Content filtering",
        "Online": "yes"
    }
}
```


Listing 2.4. ToT framework with few-shot learning and CoT classification results of 17 incidents

E Zero-Shot Learning Classification Results

The zero-shot learning classification generally got the geolocation fields and affected population information without much issue.


```
[
  {
    "id": "0",
    "Country": "Worldwide",
    "State": "",
    "City": "",
    "Continent": "Worldwide",
    "Company": "Multiple companies involved in AI development
    ",
    "Company city": "",
    "Company state": "",
    "Affected population": "Global population",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Environmental Impact
    ",
    "Subclasses": "",
    "Sub-subclass": "",
    "Area of AI Application": "General AI development and
    application",
    "Online": "No",
    "title": "AI's Environmental Footprint",
    "description": "The environmental footprint of AI,
    particularly in training large models, is significant.
    According to a study by researchers at the University of
    Massachusetts, the energy used in training certain
    popular large AI models can produce about 626,000 pounds
    of carbon dioxide. This amount is equivalent to roughly
    300 round-trip flights between New York and San Francisco
    , highlighting the substantial carbon footprint
    associated with advanced AI technologies. This data
    underscores the need for more sustainable practices in
    the field of AI to mitigate its impact on climate change
    .",
    "date occurrence": "2023-07-18",
    "date publicly known": "2023-07-18"
  },
  {
    "id": "1",
    "Country": "Worldwide",
    "State": "",
    "City": "",
    "Continent": "Worldwide",
    "Company": "YouTube",
```

```
    "Company city": "San Bruno",
    "Company state": "California",
    "Affected population": "Online Child Population",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Human Incompetence,
   Mental Health, Other",
    "Subclasses": "Administrative",
    "Sub-subclass": "",
    "Area of AI Application": "content filtering",
    "Online": "yes",
    "title": "Googles YouTube Kids App Presents
   Inappropriate Content",
    "description": "YouTubes content filtering and
   recommendation algorithms exposed children to disturbing
   and inappropriate videos.",
    "date occurrence": "2015-05-19",
    "date publicly known": "2017-03-26"
  },
  {
    "id": "2",
    "Country": "United States",
    "State": "New Jersey",
    "City": "Robbinsville",
    "Continent": "North America",
    "Company": "Amazon",
    "Company city": "Seattle",
    "Company state": "Washington",
    "Affected population": "Warehouse Workers",
    "Number of people actually affected": "24",
    "Number of people potentially affected": "54",
    "Classes of irresponsible AI use": "Human Incompetence",
    "Subclasses": "Technical",
    "Sub-subclass": "",
    "Area of AI Application": "Warehouse Automation",
    "Online": "No",
    "title": "Warehouse robot ruptures can of bear spray and
   injures workers",
    "description": "Twenty-four Amazon workers in New Jersey
   were hospitalized after a robot punctured a can of bear
   repellent spray in a warehouse.",
    "date occurrence": "2018-12-05",
    "date publicly known": "2018-12-06"
  },
  {
    "id": "3",
    "Country": "Indonesia",
    "State": "",
    "City": "Jakarta",
    "Continent": "Asia",
```

```
    "Company": "Boeing",
    "Company city": "Chicago",
    "Company state": "Illinois",
    "Affected population": "Airline Passengers, Crew Members",
    "Number of people actually affected": "189",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Human Incompetence, Other",
    "Subclasses": "Technical",
    "Sub-subclass": "",
    "Area of AI Application": "Aviation safety systems",
    "Online": "no",
    "title": "Crashes with Maneuvering Characteristics Augmentation System (MCAS)",
    "description": "A Boeing 737 crashed into the sea, killing 189 people, after faulty sensor data caused an automated manuevering system to repeatedly push the plane's nose downward.",
    "date occurrence": "2018-10-27",
    "date publicly known": "2019-03-13"
  },
  {
    "id": "4",
    "Country": "USA",
    "State": "Arizona",
    "City": "Tempe",
    "Continent": "North America",
    "Company": "Uber",
    "Company city": "San Francisco",
    "Company state": "California",
    "Affected population": "Pedestrians, Autonomous Vehicle Test Subjects",
    "Number of people actually affected": "1",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Human Incompetence, Other",
    "Subclasses": "Technical",
    "Sub-subclass": "",
    "Area of AI Application": "Autonomous Vehicles",
    "Online": "no",
    "title": "Uber AV Killed Pedestrian in Arizona",
    "description": "An Uber autonomous vehicle (AV) in autonomous mode struck and killed a pedestrian in Tempe, Arizona.",
    "date occurrence": "2018-03-18",
    "date publicly known": "2018-03-22"
  },
  {
    "id": "5",
```

```
    "Country": "United States",
    "State": "",
    "City": "",
    "Continent": "North America",
    "Company": "Not specified in the article",
    "Company city": "Not specified in the article",
    "Company state": "",
    "Affected population": "Patients undergoing robotic surgery",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Human Incompetence, Technical Difficulties",
    "Subclasses": "Technical, Malfunction, Design flaws",
    "Sub-subclass": "",
    "Area of AI Application": "Robotic surgery",
    "Online": "No",
    "title": "Collection of Robotic Surgery Malfunctions",
    "description": "Study on database reports of robotic surgery malfunctions (8,061), including those ending in injury (1,391) and death (144), between 2000 and 2013.",
    "date occurrence": "2015-07-13",
    "date publicly known": "2015-07-20"
  },
  {
    "id": "6",
    "Country": "Worldwide",
    "State": "",
    "City": "",
    "Continent": "Worldwide",
    "Company": "Microsoft",
    "Company city": "Redmond",
    "Company state": "Washington",
    "Affected population": "Twitter Users, Online Population",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Human Incompetence, Disinformation, Other",
    "Subclasses": "Technical, Textual",
    "Sub-subclass": "",
    "Area of AI Application": "Social Media Interaction",
    "Online": "yes",
    "title": "TayBot",
    "description": "Microsoft's Tay, an artificially intelligent chatbot, was released on March 23, 2016 and removed within 24 hours due to multiple racist, sexist, and anit-semitic tweets generated by the bot.",
    "date occurrence": "2016-03-24",
    "date publicly known": "2019-11-24"
```

```
  },
  {
    "id": "9",
    "Country": "United States",
    "State": "New York",
    "City": "New York City",
    "Continent": "North America",
    "Company": "New York City Department of Education",
    "Company city": "New York City",
    "Company state": "New York",
    "Affected population": "Students, Teachers",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Human Incompetence,
   Other",
    "Subclasses": "Administrative",
    "Sub-subclass": "",
    "Area of AI Application": "Educational assessment and
   teacher evaluation",
    "Online": "yes",
    "title": "NY City School Teacher Evaluation Algorithm
   Contested",
    "description": "An algorithm used to rate the
   effectiveness of school teachers in New York has resulted
    in thousands of disputes of its results.",
    "date occurrence": "2012-02-25",
    "date publicly known": "2013-10-18"
  },
  {
    "id": "10",
    "Country": "United States",
    "State": "",
    "City": "San Diego",
    "Continent": "North America",
    "Company": "Starbucks",
    "Company city": "Seattle",
    "Company state": "Washington",
    "Affected population": "Low-income single mothers, Retail
    workers, Baristas",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "130,000 (
   Starbucks baristas nationwide)",
    "Classes of irresponsible AI use": "Human Incompetence,
   Mental Health, Other",
    "Subclasses": "Technical",
    "Sub-subclass": "",
    "Area of AI Application": "Workforce management, Employee
    scheduling",
    "Online": "No",
```

```
    "title": "Kronos Scheduling Algorithm Allegedly Caused Financial Issues for Starbucks Employees",
    "description": "Kronoss scheduling algorithm and its use by Starbucks managers allegedly negatively impacted financial and scheduling stability for Starbucks employees, which disadvantaged wage workers.",
    "date occurrence": "2014-08-14",
    "date publicly known": "2015-06-02"
  },
  {
    "id": "11",
    "Country": "United States",
    "State": "Florida",
    "City": "Fort Lauderdale",
    "Continent": "North America",
    "Company": "Northpointe",
    "Company city": "Traverse City",
    "Company state": "Michigan",
    "Affected population": "African American",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Discrimination",
    "Subclasses": "Data bias, Algorithmic bias",
    "Sub-subclass": "Race, Feedback loop",
    "Area of AI Application": "Criminal justice risk assessment",
    "Online": "yes",
    "title": "Northpointe Risk Models",
    "description": "An algorithm developed by Northpointe and used in the penal system is two times more likely to incorrectly label a black person as a high-risk re-offender and is two times more likely to incorrectly label a white person as low-risk for reoffense according to a ProPublica review.",
    "date occurrence": "2016-05-23",
    "date publicly known": "2016-05-22"
  },
  {
    "id": "13",
    "Country": "Worldwide",
    "State": "",
    "City": "",
    "Continent": "Worldwide",
    "Company": "Google's Alphabet",
    "Company city": "Mountain View",
    "Company state": "California",
    "Affected population": "Online Population",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
```

```
    "Classes of irresponsible AI use": "Discrimination, Human Incompetence",
    "Subclasses": "Data bias, Algorithmic bias, Technical",
    "Sub-subclass": "Race, Sexual Orientation, Other, Feedback loop, Other",
    "Area of AI Application": "content filtering, online discussion moderation",
    "Online": "yes",
    "title": "High-Toxicity Assessed on Text Involving Women and Minority Groups",
    "description": "Google's Perspective API, which assigns a toxicity score to online text, seems to award higher toxicity scores to content involving non-white, male, Christian, heterosexual phrases.",
    "date occurrence": "2017-02-27",
    "date publicly known": "2021-02-09"
  },
  {
    "id": "14",
    "Country": "Worldwide",
    "State": "",
    "City": "",
    "Continent": "Worldwide",
    "Company": "Google",
    "Company city": "Mountain View",
    "Company state": "California",
    "Affected population": "Ethnic and religious minorities, LGBTQ community",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Discrimination",
    "Subclasses": "Data bias",
    "Sub-subclass": "Race, Sexual Orientation",
    "Area of AI Application": "Sentiment analysis, natural language processing",
    "Online": "yes",
    "title": "Biased Sentiment Analysis",
    "description": "Google Cloud's Natural Language API provided racist, homophobic, amd antisemitic sentiment analyses.",
    "date occurrence": "2017-10-26",
    "date publicly known": "2017-10-25"
  },
  {
    "id": "167",
    "Country": "United States",
    "State": "California",
    "City": "Stanford",
    "Continent": "North America",
    "Company": "Stanford Graduate School of Business",
```

```
    "Company city": "Stanford",
    "Company state": "California",
    "Affected population": "LGBTQ",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Discrimination,
   Pseudoscience, Mental Health",
    "Subclasses": "Data bias, Facial",
    "Sub-subclass": "Sexual Orientation",
    "Area of AI Application": "Facial recognition analysis",
    "Online": "yes",
    "title": "Researchers' Homosexual-Men Detection Model
   Denounced as a Threat to LGBTQ Peoples Safety and
   Privacy",
    "description": "Researchers at Stanford Graduate School
   of Business developed a model that determined, on a
   binary scale, whether someone was homosexual using only
   his facial image, which advocacy groups such as GLAAD and
    the Human Rights Campaign denounced as flawed science
   and threatening to LGBTQ folks.",
    "date occurrence": "2017-09-07",
    "date publicly known": "2017-10-08"
  },
  {
    "id": "382",
    "Country": "United Kingdom",
    "State": "",
    "City": "London",
    "Continent": "Europe",
    "Company": "Meta",
    "Company city": "Menlo Park",
    "Company state": "California",
    "Affected population": "Online Female Population",
    "Number of people actually affected": "1",
    "Number of people potentially affected": "Unknown",
    "Classes of irresponsible AI use": "Mental Health,
   Disinformation",
    "Subclasses": "Textual",
    "Sub-subclass": "",
    "Area of AI Application": "content filtering",
    "Online": "yes",
    "title": "Instagram's Exposure of Harmful Content
   Contributed to Teenage Girls Suicide",
    "description": "Instagram was ruled by a judge to have
   contributed to the death of a teenage girl in the UK
   allegedly through its exposure and recommendation of
   suicide, self-harm, and depressive content.",
    "date occurrence": "2017-11-21",
    "date publicly known": "2022-09-30"
  },
```

```
{
  "id": "451",
  "Country": "United Kingdom",
  "State": "",
  "City": "London",
  "Continent": "Europe",
  "Company": "Stability AI",
  "Company city": "London",
  "Company state": "",
  "Affected population": "Artists, Photographers, Content Creators",
  "Number of people actually affected": "Unknown",
  "Number of people potentially affected": "Thousandspossibly millions",
  "Classes of irresponsible AI use": "Copyright Violation",
  "Subclasses": "",
  "Sub-subclass": "",
  "Area of AI Application": "AI art tool, Image generation",
  "Online": "yes",
  "title": "Stable Diffusion's Training Data Contained Copyrighted Images",
  "description": "Stability AI reportedly scraped copyrighted images by Getty Images to be used as training data for Stable Diffusion model.",
  "date occurrence": "2022-10-16",
  "date publicly known": "2023-01-16"
},
{
  "id": "505",
  "Country": "Belgium",
  "State": "",
  "City": "",
  "Continent": "Europe",
  "Company": "Chai Research",
  "Company city": "Silicon Valley",
  "Company state": "California",
  "Affected population": "Chatbot Users",
  "Number of people actually affected": "1",
  "Number of people potentially affected": "Unknown",
  "Classes of irresponsible AI use": "Mental Health, Human Incompetence, Other",
  "Subclasses": "Technical",
  "Sub-subclass": "",
  "Area of AI Application": "Conversational AI",
  "Online": "yes",
  "title": "Man Reportedly Committed Suicide Following Conversation with Chai Chatbot",
  "description": "A Belgian man reportedly committed suicide following a conversation with Eliza, a language
```

```
    model developed by Chai that encouraged the man to commit
     suicide to improve the health of the planet.",
    "date occurrence": "2023-03-27",
    "date publicly known": "2023-03-27"
  },
  {
    "id": "39",
    "Country": "Worldwide",
    "State": "",
    "City": "",
    "Continent": "Worldwide",
    "Company": "Multiple companies including Adobe,
    University of Washington, Lyrebird, and others mentioned
    in the context of developing AI technology",
    "Company city": "Multiple locations",
    "Company state": "",
    "Affected population": "General Public",
    "Number of people actually affected": "Unknown",
    "Number of people potentially affected": "Worldwide
    population with internet access",
    "Classes of irresponsible AI use": "Disinformation,
    Mental Health, Other",
    "Subclasses": "Textual, Image, Video, Audio",
    "Sub-subclass": "",
    "Area of AI Application": "Media and information
    dissemination, including video conferencing, virtual
    reality, and potentially any form of digital
    communication",
    "Online": "yes",
    "title": "Deepfake Obama Introduction of Deepfakes",
    "description": "University of Washington researchers made
     a deepfake of Obama, followed by Jordan Peele",
    "date occurrence": "2017-07-01",
    "date publicly known": "2017-07-18"
  }
]
```


Listing 2.5. Zero-Shot Learning Classification Results of 17 incidents

F Table comparison of classification results

Table 1. Comparison of Classification Results for Incident 13 (Part 1)

Method	ID	Country	State	City	Continent	Company
Manual	13	United States	-	-	North America	Google LLC
Zero-Shot	13	Worldwide	-	-	Worldwide	Google's Alphabet
Few-Shot CoT	13	Worldwide	-	-	Worldwide	Google
ToT Few-Shot and CoT	13	United States	-	-	North America	Google's Alphabet

Table 2. Comparison of Classification Results for Incident 13 (Part 2)

Co. City	Co. State	Affected Pop.	Classes	Subclasses
Mountain View	California	Minorities, women, etc.	Discrimination	Data bias
Mountain View	California	Online Population	Discrimination, Human Incompetence	Data bias, Algorithmic bias, Technical
Mountain View	California	Online Users	Discrimination, Human Incompetence	Data bias, Algorithmic bias, Technical
Mountain View	California	Online Users	Discrimination, Human Incompetence, Disinformation	Data bias, Algorithmic bias, Technical, Textual

Table 3. Comparison of Classification Results for Incident 4 (Part 1)

Method	ID	Country	State	City	Continent	Company
Manual	4	United States	Arizona	Tempe	North America	Uber
Zero-Shot	4	USA	Arizona	Tempe	North America	Uber
Few-Shot CoT	4	USA	Arizona	Tempe	North America	Uber
ToT Few-Shot	4	USA	Arizona	Tempe	North America	Uber

Table 4. Comparison of Classification Results for Incident 4 (Part 2)

Co. City	**Co. State**	**Affected Pop.**	**Classes**	**Subclasses**	**AI Application**	**Online**
San Francisco	California	Pedestrians	Human incompetence	Technical	Autonomous driving	No
San Francisco	California	Pedestrians, AV Test Subjects	Human Incompetence, Other	Technical	Autonomous Vehicles	No
San Francisco	California	Pedestrians, AV Users, Public	Human Incompetence, Other	Technical	Autonomous Vehicles	No
San Francisco	California	Pedestrians, AV Test Subjects	Human Incompetence, Other	Technical	Autonomous Vehicles	No