MIT AI Risk Initiative

MIT FutureTech
Innovations that Shape the World

AI Incident Tracker June 2026 Update

Measuring and Improving the Reliability of LLM Incident Classifications: A Pilot Validation Study

June 2026

Authors

William Fowler, Simon Mylius and Peter Slattery

Executive Summary

The AI Incident Tracker uses large language models (LLMs) to process incident reports from the AI Incident Database (AIID) and classify them along five taxonomies, producing standardized data that policymakers and researchers can use to track trends. Because the tracker relies on automated LLM classification rather than manual review, we need to understand whether LLMs can classify incidents as consistently as expert humans. This report describes a pilot study to compare the performance of eight LLMs against expert human reviewers and identify opportunities to improve the incident tracker pipeline.

What We Did

- We asked two expert reviewers to independently read reports of 10 incidents and classify each one using five taxonomies: Harm Severity (how serious the harm was, rated across 10 harm-types), EU AI Act Risk Level (the regulatory risk tier under the EU AI Act), Causal Taxonomy (what stage of the AI lifecycle the harm stems from and whether it was intentional), Domain Taxonomy (7 high-level categories of risks, such as discrimination or malicious use), and Subdomain Taxonomy (24 more specific subcategories within domains).
- We compared the reviewers' independent classifications to measure human-human agreement. This provided a human reliability baseline, showing how consistently expert reviewers agreed before discussion.
- After completing their independent reviews, the reviewers resolved disagreements and produced consensus labels for each incident. We used these consensus labels and human reliability baseline for evaluating the performance of seven models against the model currently used by the tracker.
- Where LLM performance fell substantially below the human-human agreement baseline, we iteratively refined the LLM prompt until the agreement reached this baseline level.

What We Found

- Opus 4.6 and Kimi K2.5 were the two strongest models in our evaluation and met or exceeded the currently used model, Sonnet 4, on most taxonomies. Opus 4.6 ranked highest for self-consistency (how frequently the model produced the same classification given the same input).
- On three of the five taxonomies (Harm Severity, Domain, and Causal), several frontier models met or exceeded the human-human agreement baseline without any modifications to prompts.
- On Subdomain, frontier models sat just below the human baseline, by a margin too small to read as meaningful at this sample size. On the EU AI Act Risk Level taxonomy they performed substantially worse.
- After targeted prompt refinement, Opus 4.6 matched or exceeded the human-human agreement baseline on all five taxonomies on this pilot sample.

Introduction

Incident reporting has contributed to major safety improvements in high-stakes sectors such as aviation, healthcare, and nuclear power by helping industry and regulators identify recurring risks, improve practices, and prevent future harm. As AI systems become more capable and widely deployed, post-deployment incident monitoring is increasingly important for understanding emerging risks and prioritizing mitigations.

Several databases already collect reports of AI-related incidents, including the AI Incident Database, the OECD's AI Incident Monitor, the AIAAIC Repository, the Political Deepfakes Database, and others. The MIT AI Incident Tracker project builds on these efforts by classifying each incident against a set of risk taxonomies and rating the severity of the harm involved. This structure makes it easier to identify patterns across incidents, compare reports from different sources, and communicate trends in ways that are useful to policymakers and other stakeholders.

The current version of the tracker applies several taxonomies to classify incidents from the AI Incident Database in a consistent format. The results are presented through interactive visualizations, allowing users to explore patterns in AI incidents over time, compare incidents across MIT risk taxonomies, examine the severity of associated harms, and identify broader trends in reported AI-related failures. These outputs have been used to report on incident trends, including in the 2026 World Disasters Report and Time magazine.

The Incident Tracker uses large language models (LLMs) to make incident classification scalable. As the number of reported AI incidents grows, human review becomes increasingly impractical given the volume of reports and the time required for careful classification.

However, LLMs can make mistakes when interpreting incident reports and assigning classifications. For instance, they can occasionally hallucinate details and can exhibit a score-range bias, where ratings cluster within a narrow band of an ordinal scale rather than using the full range.

Because the tracker relies on automated LLM classification rather than manual review, we need to understand where LLMs can classify incidents as consistently as expert humans and where their judgments are less reliable. This pilot study therefore explores three questions:

1. How do LLM ratings compare to expert human consensus in classifying AI incidents?
2. Which taxonomies pose particular difficulty for LLM ratings?
3. What changes can we make to improve the classification pipeline?

By answering these questions, we aim to further refine the Incident Tracker as a proof-of-concept for scalable incident analysis and to lay the groundwork for a larger validation study.

Methodology

To assess whether our LLM classifications are comparable to how human reviewers would classify incidents, we conducted a small pilot validation study.

Determining the human baseline and consensus

We selected 10 incident reports from the AI Incident Database. The sample included one incident from each of the seven MIT AI Risk Domain categories, plus one additional incident from each of the three most common domains in the database: Discrimination & Toxicity, Malicious Actors, and AI System Safety, Failures, & Limitations.

For each incident, two reviewers with backgrounds in AI research independently read the raw AI Incident Database reports and completed a classification survey. Reviewers applied the same criteria and instructions used by the LLM pipeline. They were instructed not to use external information or prior knowledge about the incident, and to base their classifications only on the evidence contained in the raw reports.

After the independent reviews were complete, the two reviewers discussed their classifications and resolved disagreements to produce a consensus label for each rating. We used these consensus labels as the benchmark for evaluating LLM performance.

Model Selection

We selected a mix of frontier models and lower-cost options:

- Haiku 4.5
- Sonnet 4 (current)
- Sonnet 4.6
- Opus 4.6
- GPT 5.2
- Gemini 3 Flash Preview
- Gemma 3 27B
- Kimi K2.5

Evaluation

We used two metrics to measure agreement between any two raters (whether human-human or LLM-human consensus):

Exact Match Percentage: the fraction of incidents on which the two raters chose the same category.

Quadratic Weighted Kappa (QWK): a standard inter-rater agreement metric for taxonomies where the categories have a natural order, such as Harm Severity (low to high) and EU AI Act Risk Level (minimal to unacceptable).

QWK rewards close-but-not-exact agreement (e.g., rating an incident "2" when the other rater chose "3") and penalizes large disagreements (e.g., "1" vs "4") more heavily. QWK adjusts for the rate of agreement expected by chance, producing a score between -1 and 1, where 1 indicates perfect agreement, 0 indicates the same level of agreement as would be expected if scores were assigned randomly, and negative scores indicate less agreement than there would be by chance.

QWK agreement is usually interpreted as: <0 Poor, 0.00–0.20 Slight, 0.21–0.40 Fair, 0.41–0.60 Moderate, 0.61–0.80 Substantial, and 0.81–1.00 Almost Perfect.

We used Exact Match Percentage as the primary metric for taxonomies where the categories have no natural order (Causal, Domain, and Subdomain). We used QWK in addition to Exact Match for ordinal taxonomies because Exact Match alone treats every disagreement as equally important, which is misleading when the categories are ordered.

Beyond agreement with the human baseline, we also measured each model's consistency: how often it produces the same rating when run three times on the same incident. Higher self-consistency reduces the variance in incident classifications.

Iteration process

We used an iterative evaluation process. We first ran each model using the original instruction prompt and measured agreement with human consensus on 10 incidents. If the tested LLMs consistently performed substantially below the human-human baseline, we revised the instruction prompt and re-ran the evaluation.

Figure 1 summarizes our high-level end-to-end evaluation and iteration process. The diagram shows how human consensus labels and LLM ratings feed into the same metric computation, and how taxonomies that fall below the baseline trigger a prompt-revision loop.

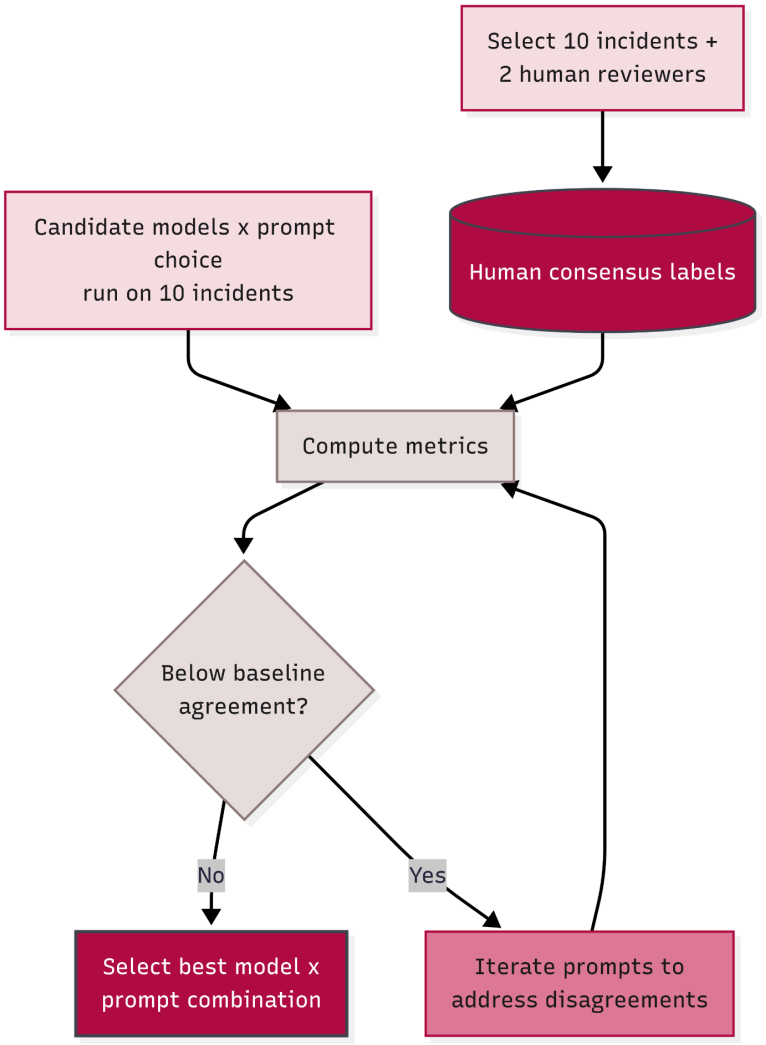

Figure 1. *Methodology for evaluating and improving LLM classification of AI incidents.*

Results

Evaluating the human baseline

First, to establish a baseline human reviewer performance, we calculated the agreement scores for individual human reviewers against each other. Since our evaluation includes both ordinal taxonomies (Harm Severity ratings and EU AI Act scores) as well as categorical ratings (Causal, Domain, and Subdomain taxonomies), we show QWK for the ordinal scores and an exact match percentage for all. By the Landis & Koch interpretation scale, the human-human agreement is itself only "Fair" for Harm Severity (0.31) and "Moderate" for EU AI Act Risk Level (0.56).

The fact that human reviewers disagreed on some classifications, is not surprising; some disagreement is normal in reviews. In this case, several factors likely contributed, including differences in reviewer experience, incomplete incident reports, and taxonomies that require judgment about severity, causality, regulatory context, and risk category.

Table 1. Human-Human Agreement Scores.

Taxonomy	QWK (where applicable)	Exact Match %
Harm Severity* (N=100)	0.31	57%
EU AI Act (N=10)	0.56	60%
Causal* (N=30)	N/A	50%
Domain (N=10)	N/A	80%
Subdomain (N=10)	N/A	80%

**Harm Severity has N=100 because each incident receives 10 Harm Severity sub-ratings; Causal Taxonomy has N=30 because each incident receives 3 causal sub-ratings; the other taxonomies have one rating per incident (N=10). The N=10 taxonomies (EU AI Act, Domain, Subdomain) carry substantial uncertainty: on an N=10 exact-match scale a single reclassified incident shifts the score by 0.10, and QWK is more sensitive still. We therefore base model selection on patterns that hold across taxonomies rather than on small gaps between adjacent models, and treat differences of one to two incidents as within noise.*

Initial Results

Figures 2 and 3 show LLM agreement with human consensus using the original instruction prompt. Each chart includes the H1 vs H2 human baseline for reference.

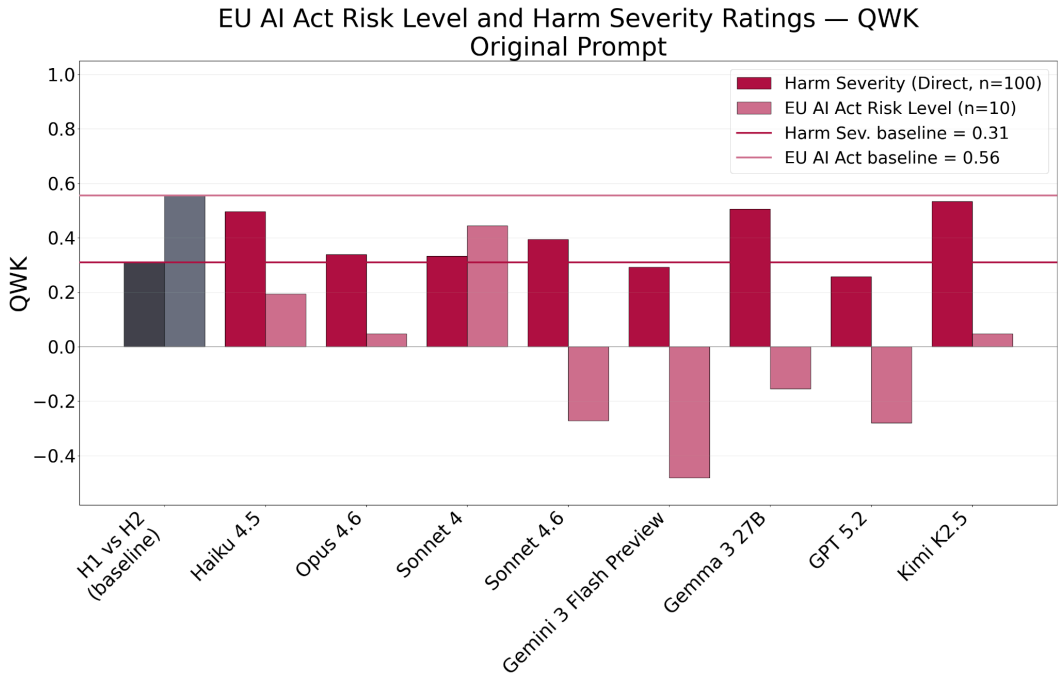

Figure 2. *QWK agreement against human consensus for the two ordinal taxonomies (Harm Severity and EU AI Act Risk Level) under the original prompt. Lines mark the human-human baseline for each taxonomy.*

These results surfaced the EU AI Act Risk Level as the clearest weak-spot: four of the eight models (Sonnet 4.6, Gemini 3 Flash Preview, Gemma 3 27B, and GPT 5.2) scored a negative QWK, meaning their ratings agreed with human consensus less than would be expected by

random chance (Figure 2). The categorical taxonomies (Domain, Subdomain, and Causal) showed tighter agreement, with no model falling far below the human-human baseline. Every model cleared the Causal baseline comfortably; on Domain, most clustered around the baseline with only Gemini 3 Flash Preview clearly above it; and on Subdomain, those that fell short did so by about 10 percentage points (one incident) which we do not consider meaningful for the N=10 taxonomies (Figure 3).

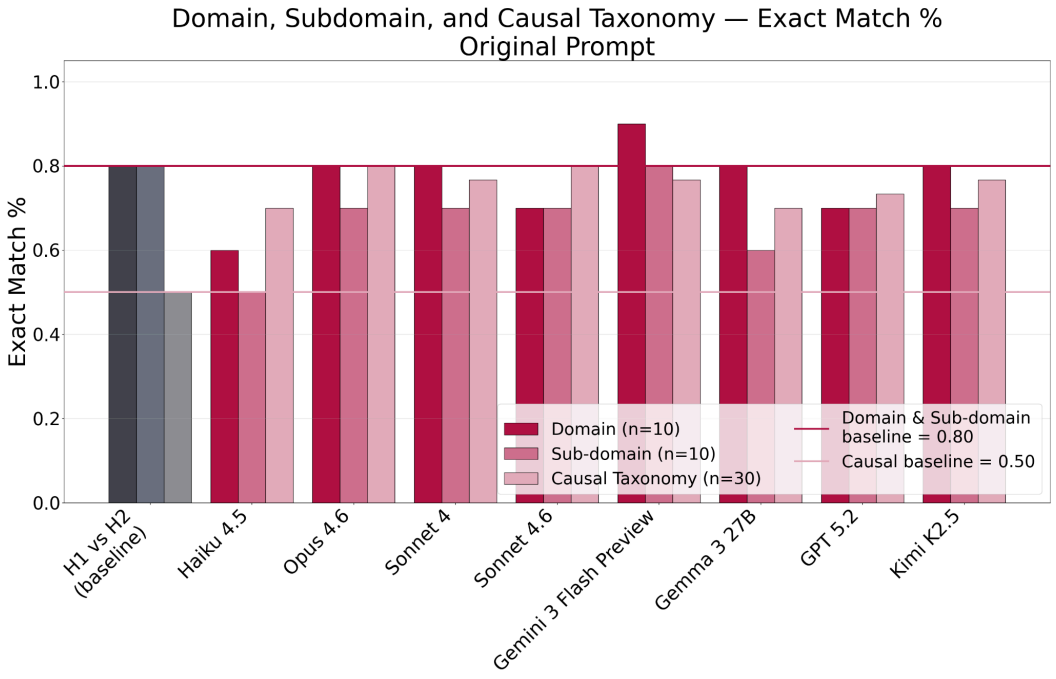

Figure 3. *Exact-match agreement against human consensus for the three categorical taxonomies (left-to-right: Domain, Subdomain, and Causal) under the original prompt. Human-human baselines are shown for each taxonomy.*

Figures 2 and 3 also point to the strongest candidate models. Haiku 4.5, Gemma 3 27B, and Kimi K2.5 cleared the Harm Severity baseline most decisively, while Opus 4.6 and Kimi K2.5 were the only models to meet or exceed Sonnet 4 (the current production model) on nearly every taxonomy, identifying them as our strongest candidates to replace it.

EU AI Act Prompt Revisions

We then implemented five different clarifications to the EU AI Act risk level prompt. We asked Claude Opus 4.6 to read through each model's reasoning for its EU AI Act ratings, paying particular attention to ratings that disagreed with human consensus, and to propose five distinct clarifications to the prompt to address them. We tested each individually, and compared them to a sixth prompt combining all five clarifications.

Opus 4.6 with the combined prompt scored highest overall, matching or exceeding human baseline agreement across all five taxonomies (Figure 4).

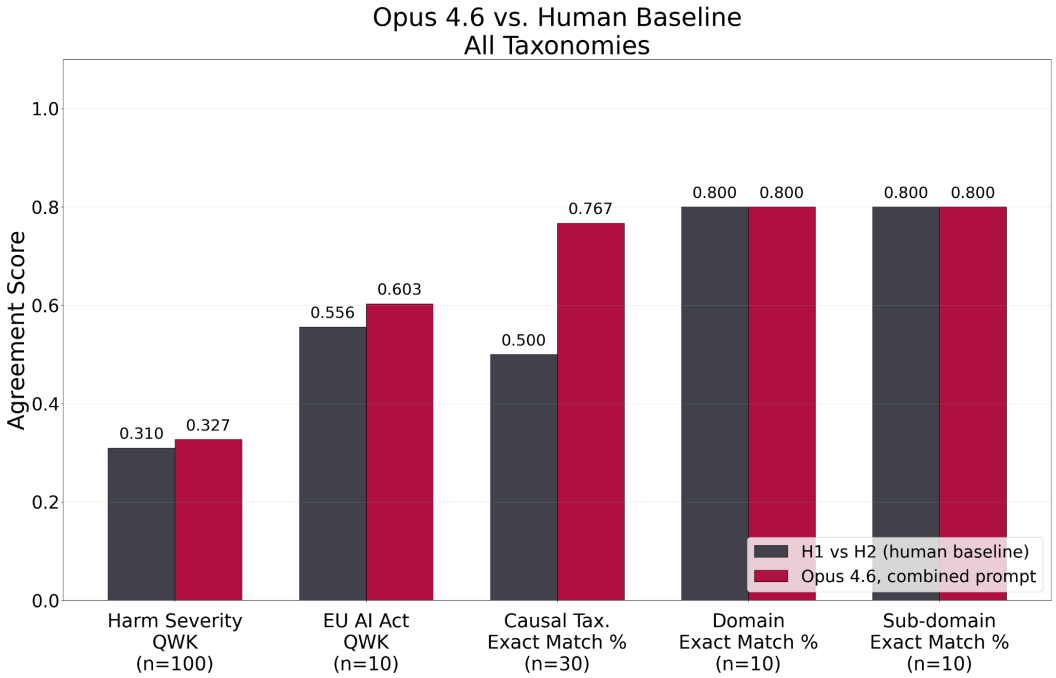

Figure 4. *Opus 4.6 with the combined EU AI Act clarifications prompt compared to the human-human baseline across all five taxonomies.*

Consistency

Opus 4.6 ranked highest for self-consistency among the models tested, achieving a mean of 94% across the five taxonomies evaluated. Haiku 4.5 had the lowest average self-consistency, at 69%.

Future Work

Our study had several limitations that future work could address. The human consensus labels used as the benchmark for evaluating LLM performance were based on a small review sample: two reviewers per incident across 10 incidents. Expanding the number of incidents would improve the precision of model-performance estimates, while increasing the number of reviewers per incident would strengthen the reliability of the human consensus labels. Such a study might also create value by producing a more substantial labeled dataset which could be used to fine-tune an LLM specifically for incident classification.

While we required models to select a single category within each taxonomy, future work could instead apply relevance scores to each possible category. For example, instead of selecting one of the seven domain labels for an incident, the model would score the relevance of all seven labels. This would provide a more granular representation of the model's classification judgments and provide richer insights into incidents where more than one risk domain or subdomain apply.

Conclusion

We ran a human-review validation study to evaluate the Incident Tracker's LLM-based classification pipeline against the consensus of expert human reviewers. The evaluation covered five taxonomies: Harm Severity, EU AI Act Risk Level, Causal Taxonomy, Domain Taxonomy, and Subdomain Taxonomy. Together, these taxonomies capture how serious the harm was, how the incident maps to regulatory risk categories, what caused the harm, and which risk domain and subdomain the incident falls into.

We tested seven candidate models, Haiku 4.5, Sonnet 4.6, Opus 4.6, GPT 5.2, Gemini 3 Flash Preview, Gemma 3 27B, and Kimi K2.5, against Sonnet 4, the model currently used by the tracker. We then iteratively refined the written instructions given to the LLM, focusing on EU AI Act Risk Level, the taxonomy where model performance most clearly fell below the human-human baseline.

How do LLM ratings compare to expert human consensus in classifying AI incidents?

Overall, the results suggest that LLM-based incident classification can approach the reliability of expert human review. Some frontier models (Opus 4.6 and Kimi K2.5) met or exceeded human baseline in three of the five taxonomies (Harm Severity, Domain, and Causal) without any modifications to prompts. For the EU AI Act Risk Level taxonomy, prompt clarifications were necessary to reach the human baseline. After that targeted refinement, Opus 4.6 matched or exceeded the human-human agreement baseline on all five taxonomies, on this pilot sample.
These results indicate that agreement levels comparable to the human baseline are attainable, but that model selection and prompt design are important determinants of performance.

Which taxonomies pose particular difficulty for LLM ratings?

The EU AI Act Risk Level taxonomy was the most difficult classification task for the models. It produced the lowest agreement scores across tested models, likely because the human baseline was relatively high and because models had difficulty disambiguating the definitions of the risk levels. The error pattern did not show a strong directional bias: across all model and prompt combinations tested, 43% of errors were overestimates of risk level and 57% were underestimates. Adding clarifications to the EU AI Act prompt reduced the number of errors. Domain and Subdomain sat in between, with models clustered around the human baselines. The Causal taxonomy, by contrast, was the least difficult classification task with every model clearing the human baseline comfortably.

What changes can we make to improve the classification pipeline?

Our analysis identified three main ways to improve the classification pipeline. First, we can select models based on performance on the weakest taxonomies, while ensuring that they still meet or exceed the human-human baseline on the other classification tasks. This provides a rationale for choosing models for future classification work: of the models we

tested, Opus 4.6 already met the baseline on Harm Severity, Domain, and Causal with the original prompt. It fell only marginally below baseline on Subdomain, leaving the EU AI Act as its sole clear weakness, which the prompt clarifications addressed.

Second, we can use targeted prompt refinement to improve classification performance. Rather than relying only on model selection, we can identify the taxonomies where models perform poorly, examine common sources of disagreement with human reviewers, and revise the instructions to make classification criteria clearer and easier to apply consistently.

Third, we can use another LLM to review model reasoning, identify likely causes of misclassification, and propose prompt improvements that can then be tested empirically.

Next steps

Our results suggest that LLM-based incident classification has comparable accuracy to the human baseline and is a viable approach for scaling the Incident Tracker.

The next step is to run a more comprehensive validation study based on this pilot. We will increase the sample size of incidents reviewed by human experts and further iterate targeted prompt refinement and model selection. The full set of reports from the AI Incident Database will be reclassified using the updated model selection and adapted prompts, to refresh the Incident Tracker with more reliable classifications.
Because the prompt refinements were developed using this validation sample, a set of human-reviewed incidents will be held to confirm ongoing performance as the tracker is updated.

Acknowledgments

We want to thank the following reviewers for useful contributions and feedback:

- Yan Zhu
- Lauren Nieto
- Suhani Gharia
- Spencer Michaels
- Michael Noetel
- Branwen Owen

William Fowler's work on this project was funded by the Cambridge Boston Alignment Initiative with the financial support of Coefficient Giving.