

# Automating AI Failure Tracking: Semantic Association of Reports in AI Incident Database

Diego Russo<sup>a</sup>, Gian Marco Orlando<sup>b</sup>, Valerio La Gatta<sup>c</sup> and Vincenzo Moscato<sup>b</sup>

<sup>a</sup>University of Bergamo, Department of Management, Information and Production Engineering, Via Pasubio 7b, Dalmine (BG), 24044, Italy

<sup>b</sup>University of Naples Federico II, Department of Electrical Engineering and Information Technology, via Claudio 21, 80125, Naples, Italy

<sup>c</sup>Northwestern University, Department of Computer Science, McCormick School of Engineering and Applied Science

ORCID (Diego Russo): <https://orcid.org/0009-0007-1095-5168>, ORCID (Gian Marco Orlando): <https://orcid.org/0009-0004-7136-1804>, ORCID (Valerio La Gatta): <https://orcid.org/0000-0002-5941-4684>, ORCID (Vincenzo Moscato): <https://orcid.org/0000-0002-0754-7696>

## Abstract.

Artificial Intelligence (AI) systems are transforming critical sectors such as healthcare, finance, and transportation, enhancing operational efficiency and decision-making processes. However, their deployment in high-stakes domains has exposed vulnerabilities that can result in significant societal harm. To systematically study and mitigate these risks, initiatives like the AI Incident Database (AIID) have emerged, cataloging over 3,000 real-world AI failure reports. Currently, associating a new report with the appropriate *AI Incident* relies on manual expert intervention, limiting scalability and delaying the identification of emerging failure patterns.

To address this limitation, we propose a retrieval-based framework that automates the association of new reports with existing *AI Incidents* through semantic similarity modeling. We formalize the task as a ranking problem, where each report—comprising a title and a full textual description—is compared to previously documented *AI Incidents* based on embedding cosine similarity. Benchmarking traditional lexical methods, cross-encoder architectures, and transformer-based sentence embedding models, we find that the latter consistently achieve superior performance. Our analysis further shows that combining titles and descriptions yields substantial improvements in ranking accuracy compared to using titles alone. Moreover, retrieval performance remains stable across variations in description length, highlighting the robustness of the framework. Finally, we find that retrieval performance consistently improves as the training set expands. Our approach provides a scalable and efficient solution for supporting the maintenance of the AIID.

## 1 Introduction

Artificial Intelligence (AI) systems are increasingly deployed across high-stakes domains, where their decisions significantly impact human lives and societal structures. In healthcare, they have improved diagnostic accuracy and accelerated treatment workflows [6, 7, 43]. Financial institutions leverage AI for credit scoring, fraud detection, and algorithmic trading, achieving gains in efficiency and risk management [1, 3, 8, 11]. AI also underpins key capabilities in au-

tonomous transportation [10], criminal justice systems [34], large-scale content moderation [14], and military operations, where it supports surveillance, target recognition, and mission planning [18].

However, real-world deployments have shown that AI systems can fail in unpredictable, opaque, and sometimes harmful ways, especially when deployed at scale. For instance, the use of machine learning to support clinical decision-making has been shown to exacerbate existing health disparities, with certain models underperforming for specific demographic groups [31]. Similarly, in the financial sector, AI-driven credit scoring systems have inadvertently resulted in discriminatory outcomes against minority populations [16]. Additional failures span from accidents involving autonomous driving systems<sup>1</sup> to wrongful arrests caused by facial recognition misidentifications<sup>2</sup>, emphasizing the potential of AI to reinforce systemic biases, including racism<sup>3</sup> and misogyny<sup>4</sup>.

Given the growing evidence of AI failures across diverse application domains, there is a critical need to systematically document and study these incidents to prevent their recurrence and support the development of safer, more reliable systems. Similar to long-established safety infrastructures in aviation and cybersecurity (e.g., FAA, NASA ASRS, CVE), the AI Incident Database (AIID) was established as a structured repository of real-world failures involving AI technologies [23]. By cataloging detailed incident reports, AIID promotes transparency, accountability, and continuous improvement in AI development. Maintained by a coalition of academic, industrial, and non-profit stakeholders, the database currently hosts over 3,000 curated reports, enabling systematic investigation of failure patterns with the goal of reducing the likelihood that similar failures recur. At the core of the AIID is the concept of *AI Incident*, an alleged harm or near-harm event in which an AI system is implicated [24]. Each incident is typically described through one or more reports — narrative accounts derived from journalistic sources, academic literature, or in-

<sup>1</sup> <https://incidentdatabase.ai/cite/23/>

<sup>2</sup> <https://incidentdatabase.ai/cite/74/>

<sup>3</sup> <https://incidentdatabase.ai/cite/60/>

<sup>4</sup> <https://incidentdatabase.ai/cite/47/>

stitutional investigations — which provide details about the context, causes, consequences, and actors involved. These reports consist of a title and a free-text description, both of which can vary considerably in length and level of detail, and are complemented by structured metadata such as system functionality, affected stakeholders, and harm types.

However, when a new report is submitted to the AIID, its association with an existing *AI Incident* is still handled manually by human editors [30]. This reliance on manual expert intervention poses a scalability bottleneck as the volume of incident reports continues to grow, and it introduces potential inconsistencies due to subjective judgment in the classification process. Moreover, the lack of automation limits the ability to rapidly analyze emerging failure patterns across related incidents.

In this work, we address this gap by proposing a method to assist AIID curators in the task of automatically linking newly submitted reports to existing *AI Incidents*. Specifically, we frame the problem as a retrieval task: the textual content of each report — including its title and description — is compared to previously recorded *AI Incidents* through semantic similarity-based ranking. The goal is to rank historical incidents based on their semantic relevance to the new report. To assess the robustness and generalizability of our approach, we systematically evaluate ranking performance under a variety of conditions — including different input representations (e.g., title only vs. title and description), variations in description length, and temporal dynamics. To operationalize our investigation, we formulate the following research questions (RQs):

- RQ1:** *Does leveraging both the report’s title and description improve retrieval performance?*
- RQ2:** *Does the length of a report’s description affect the robustness of retrieval performance?*
- RQ3:** *How does retrieval performance change as the size of the training data increases?*

To answer these questions, we first conduct comprehensive experiments using both traditional lexical baselines and modern transformer-based sentence embedding models. Our findings demonstrate that sentence transformer models consistently outperform all other approaches. After identifying the best-performing model (*multi-qa-MiniLM-L6-cos-v1*), we systematically addressed the three RQs.

First, we find that while titles alone carry meaningful semantic information, incorporating full descriptions significantly improves retrieval performance, with an average gain of 15–25 percentage points in ranking metrics. Second, our analysis shows that shorter descriptions slightly outperform longer ones, although the performance differences are minimal, suggesting that our system is robust to variations in description length. Finally, we show that retrieval performance steadily improves as the size of the training data increases. These findings highlight the feasibility and robustness of automating the linking of new reports to existing *AI Incidents*, paving the way for more scalable and consistent maintenance of the AI Incident Database.

## 2 Related Works

This section reviews prior work relevant to our approach. We first examine advances in retrieval models. We then discuss how structured reporting systems have supported safety practices in fields like aviation and cybersecurity. Finally, we turn to the AI Incident Database

itself, highlighting its growing importance and the need for scalable tools to support its maintenance.

### 2.1 Semantic Retrieval Models

Retrieving semantically relevant documents from large-scale textual corpora is a fundamental task in information retrieval. Traditional approaches have long relied on term-based probabilistic models such as BM25 [35], which offer competitive performance and scalability. However, these methods are inherently limited by their reliance on lexical overlap, making them vulnerable to the vocabulary mismatch problem [13] and unable to capture deeper semantic relations. Earlier attempts to mitigate this limitation employed term dependency and topic modeling techniques [25, 4, 20], though these approaches proved insufficient for more nuanced semantic reasoning. More recent improvements, such as retrieval pipelines combining BM25 with T5-based document or query generation [28], seek to enrich input representations through generative query expansion.

With the advent of transformer-based language models, neural approaches to retrieval have gained prominence, enabling more expressive semantic matching through learned text representations. These methods are generally classified into sparse and dense retrieval paradigms [15]. Sparse retrieval models improve upon classical term-based methods by enhancing token-level representations while maintaining a sparse indexing structure (e.g., DeepCT [9], docT5query [27]). Dense retrieval methods, on the other hand, map both queries and documents into a continuous embedding space using dual-encoder architectures, computing similarity scores via a pre-defined function  $f$ . Depending on the granularity of representation, dense models can be further divided into term-level approaches — such as DC-BERT [26] and ColBERT [19] — which preserve fine-grained token interactions, and document-level approaches — including Sentence-BERT [32], DPR [17], and the MultiQA family [41] — which yield global embeddings for entire texts.

In this work, we evaluate a diverse set of retrievers in the context of the AI Incident Database. Our objective is to understand which models are most effective for linking newly submitted reports to semantically related historical *AI Incidents*, where lexical overlap is often minimal and semantic cues are subtle yet crucial for accurate matching.

### 2.2 Incident Databases in Safety-Critical Domains

Several safety-critical industries have long-established incident reporting systems that serve as foundational infrastructure for risk mitigation and continuous improvement. For instance, in the aviation sector, databases maintained by agencies such as the Federal Aviation Administration (FAA)<sup>5</sup> and the National Aeronautics and Space Administration (NASA)<sup>6</sup> have played a pivotal role in advancing operational safety standards [38]. These repositories systematically document accidents—defined as events resulting in serious damage or casualties—and incidents, which are precursors or near-misses that expose vulnerabilities within existing processes. Through the aggregation of structured event data, technical reports, and expert analyses, these systems have supported the development of robust safety protocols and informed regulatory decision-making.

A similar model exists in the domain of cybersecurity, where the Common Vulnerabilities and Exposures<sup>7</sup> (CVE) system provides a

<sup>5</sup> [https://www.faa.gov/data\\_research/accident\\_incident](https://www.faa.gov/data_research/accident_incident)

<sup>6</sup> <https://asrs.arc.nasa.gov/>

<sup>7</sup> <https://cve.mitre.org/>

standardized nomenclature for publicly disclosed software vulnerabilities. Maintained by The MITRE Corporation, the CVE framework enables consistent identification, cataloging, and dissemination of security issues across heterogeneous environments.

In safety-critical applications where AI failures can lead to significant harm, the AI Incident Database<sup>8</sup> has recently emerged as a key resource. By systematically cataloging real-world incidents involving AI across diverse sectors, the AIID enables the identification of recurring failure modes and risk factors—critical insights for developing robust, evidence-based mitigation strategies.

### 2.3 The AI Incident Database

The AI Incident Database is a structured repository of documented failures, harms, and unintended consequences arising from the deployment of AI systems [23]. It aims to support transparency, accountability, and risk analysis in AI development by cataloging real-world incidents reported across media sources, academic literature, and governmental investigations. Each record — referred to as an *AI Incident* — denotes a case in which an AI system is implicated in causing, or nearly causing, harm. Incidents are described with rich metadata, including the organizations involved, system functionality, affected stakeholders, types of harm, and may be linked to multiple reports over time. To capture recurring patterns of failure, the AIID introduces the concept of an *AI Incident Variant*, defined as an event that shares similar causes, harms, and system-level characteristics with a previously reported *AI Incident*. Alongside these grounded categories, the AIID also tracks *AI Issues* — speculative or potential harms that have not yet occurred or been empirically observed but are flagged as emerging concerns in the deployment of AI technologies [24].

The AIID has been increasingly adopted as a foundational resource in diverse areas of AI research and governance, providing real-world evidence of failures and harms caused by deployed AI systems. Several recent works have leveraged AIID records to inform both normative and empirical investigations. For instance, [36] analyze the role of incident documentation in shaping governance responses to AI failures. Beyond analytical use, the AIID has also been integrated into regulatory and educational practices. [22] examined its role in shaping emerging AI regulation frameworks, while [12] evaluated its effectiveness as a pedagogical tool to build awareness of AI harms in classroom settings. The database has also been used to inform ethical risk assessments in sectors such as healthcare [5], cybersecurity [2], and automated hiring [39], emphasizing its practical relevance across domains.

Collectively, these works underscore the AIID’s growing relevance, highlighting its potential to support systematic understanding and oversight of AI harms at scale. Despite this momentum, the AIID rely on manual curation. As the volume of new reports continues to grow, there is a pressing need for scalable, automated approaches that can assist in maintaining and enriching the AIID’s structure.

In this work, we address this gap by introducing a novel retrieval-based task: automatically linking newly submitted incident reports to semantically related, pre-existing *AI Incidents*. We propose a retrieval methodology designed to assist AIID maintainers in curating the database more efficiently, enabling scalable and context-aware integration of new information.

## 3 Materials and Methods

This section first introduces the dataset used in our study and the evaluation metrics employed to assess retrieval performance. We then formalize the task of associating new reports with relevant *AI Incidents* as a semantic ranking problem and present the methodology adopted, including the design of the retrieval pipeline and its main processing stages.

### 3.1 Dataset & Metrics

We conducted our experiments using data from the AI Incident Database [23], a publicly available repository documenting real-world failures involving AI systems. The dataset comprises user-submitted, editor-approved reports that are linked to *AI Incidents*. Each report includes a title and a textual description of the event. Reports are linked to at least one corresponding incident, while a single incident may be associated with multiple reports. An illustrative example<sup>9</sup> of the relationship between an *AI Incident* and two of its associated reports is presented in Table 1.

<b>AI Incident Title</b>	Warehouse robot ruptures can of bear spray and injures workers
<b>AI Incident Description</b>	Twenty-four Amazon workers in New Jersey were hospitalized after a robot punctured a can of bear repellent spray in a warehouse.
<b>Report Title #1</b>	1 critical, 54 Amazon workers treated after bear repellent discharge in N.J. Warehouse
<b>Report Description #1</b>	A worker at an Amazon warehouse in New Jersey was in critical condition and another 54 required treatment after being exposed to bear repellent that discharged when a can was punctured by an automated machine Wednesday morning inside the building, officials said. A total of 54 workers [...] The warehouse was cleared for re-entry around 1 p.m. by the West Windsor Health Department, but an official will revisit the building before Thursday morning as a precaution.
<b>Report Title #2</b>	Amazon bear repellent accident sends 24 workers to the hospital
<b>Report Description #2</b>	On Wednesday this week 24 Amazon warehouse workers in Robbinsville New Jersey were hospitalized after a robot punctured a can of repellent according to local news reports. One employee is said to be in critical condition. The accident In all, 54 workers [...] However, Amazon hardly is at fault in this instance. An Amazon spokesperson claimed: “While any serious incident is one too many, we learn and improve our programs working to prevent future incidents.” The company claims it surveys it workers every month to gauge their perception of safety in their workplace. More about bear repellent, Robots, Amazon More news from bear repellent Robots Amazon.

**Table 1. Example of an *AI Incident* and two of its corresponding reports.** The incident describes a harmful failure involving an automated system in a warehouse setting. The two reports provide independent accounts of the same underlying event, each contributing complementary details that help contextualize and document the incident.

<sup>8</sup> <https://incidentdatabase.ai/>

<sup>9</sup> This example is drawn directly from the AI Incident Database, available at <https://incidentdatabase.ai/cite/2/>.

The dataset<sup>10</sup> comprises 815 *AI Incidents* and 3,805 reports. These records are obtained by filtering out entries labeled as *AI Issues*, which — by the AIID’s definition — refer to alleged harms by an AI system that have yet to occur or be detected [24]. As such, they are prospective or hypothetical in nature and cannot be reliably associated with any past incident.

The performance of the proposed models is evaluated using standard ranking metrics commonly adopted in retrieval tasks: *Accuracy@K*, *Mean Reciprocal Rank (MRR@K)*, and *Normalized Discounted Cumulative Gain (NDCG@K)*, where  $K$  denotes the number of top-ranked incident predictions for a given report considered in the evaluation.

### 3.2 Methodology

In this study, we conceptualize the problem of associating newly submitted reports with existing *AI Incidents* as a retrieval task. The objective is to rank *AI Incidents* based on their semantic relevance to a given report.

The retrieval process follows a sequential pipeline, as detailed in Figure 1, which includes the following steps:

1. **Text Transformation:** For each *AI Incident* and a given report, the title and description are concatenated to form a unified textual representation for that individual entry. Subsequently, a preliminary pre-processing phase is applied, which involves standard text cleaning operations — such as the removal of non-informative characters (e.g., line breaks, emojis) — and the elimination of stopwords. The resulting texts are then converted into vector representations, according to the semantic embedding model adopted;
2. **Semantic Similarity Calculation:** The vector representation of the considered report is systematically compared against the vector representations of all existing *AI Incidents*. For each pairwise comparison, a similarity score is computed to quantify their semantic proximity;
3. **Incident Ranking:** Based on the computed similarity scores, *AI Incidents* are ranked by semantic relevance to the given report.

The final output of the pipeline is an *AI Incident List*, where existing incidents are ranked by their semantic relevance to the newly submitted report. This prioritized list is designed to assist curators in efficiently identifying the most likely matches, streamlining the process of linking new reports to documented incidents.

## 4 Experiments

In this section, we empirically evaluate the proposed retrieval framework for linking new reports to relevant *AI Incidents*. We begin by outlining the experimental setup and model configurations used in our study. We then present a comparative analysis of several retrieval models to identify the most effective approach for the task. Finally, we address the three research questions, providing a detailed analysis of the factors that influence retrieval performance.

### 4.1 Experimental Setup

To assess the performance of different semantic retrieval strategies for matching *AI Incidents* with the considered report, we conducted

<sup>10</sup> We use the official AI Incident Database snapshot available at <https://incidentdatabase.ai/research/snapshots/>. Specifically, we rely on the backup released on October 28, 2024.

experiments using both lexical and embedding-based models. Specifically, the evaluated models include BM25 [35], BM25+T5 [29], cross-encoders (i.e., *quora-roberta-base*, *ms-marco-MiniLM-L4-v2*) [21, 37], and sentence transformers (i.e., *multi-qa-distilbert-cos-v1*, *all-mpnet-base-v2*, *multi-qa-MiniLM-L6-cos-v1*) [33, 40, 42]. Cross-encoders and sentence transformers were fine-tuned on a training set derived by splitting the dataset into 75% for training, 12.5% for validation, and 12.5% for testing. All results reported in the following sections are based on the test set, which contains 475 variants. All experiments<sup>11</sup> were performed on a workstation equipped with an AMD Ryzen 7 5800H CPU (3.20 GHz), 32 GB RAM, and an NVIDIA RTX 3060 GPU.

### 4.2 Model Selection

Table 2 summarizes the performance of various models evaluated at cutoff values of  $K = 3$ ,  $K = 5$ , and  $K = 10$ . The results reveal significant distinctions among traditional lexical models, hybrid approaches, and neural embedding-based models.

Among the traditional methods, BM25 consistently exhibits the weakest performance, struggling to capture the complex semantic relationships inherent in incident-report pairs. The inclusion of synthetic queries via T5 leads to improvements in ranking accuracy, highlighting the benefits of query expansion for enhancing traditional approaches. In contrast, cross-encoder models perform sub-optimally compared to the BM25+T5 approach, failing to achieve competitive results. On the other hand, sentence transformer models consistently outperform all other approaches, demonstrating superior results across all evaluation metrics. This strong performance makes them the preferred choice for addressing the research questions in this study. Specifically, we adopt the *multi-qa-MiniLM-L6-cos-v1* model for subsequent analysis, as it consistently achieved the best performance across all values of  $K$  and evaluation metrics considered.

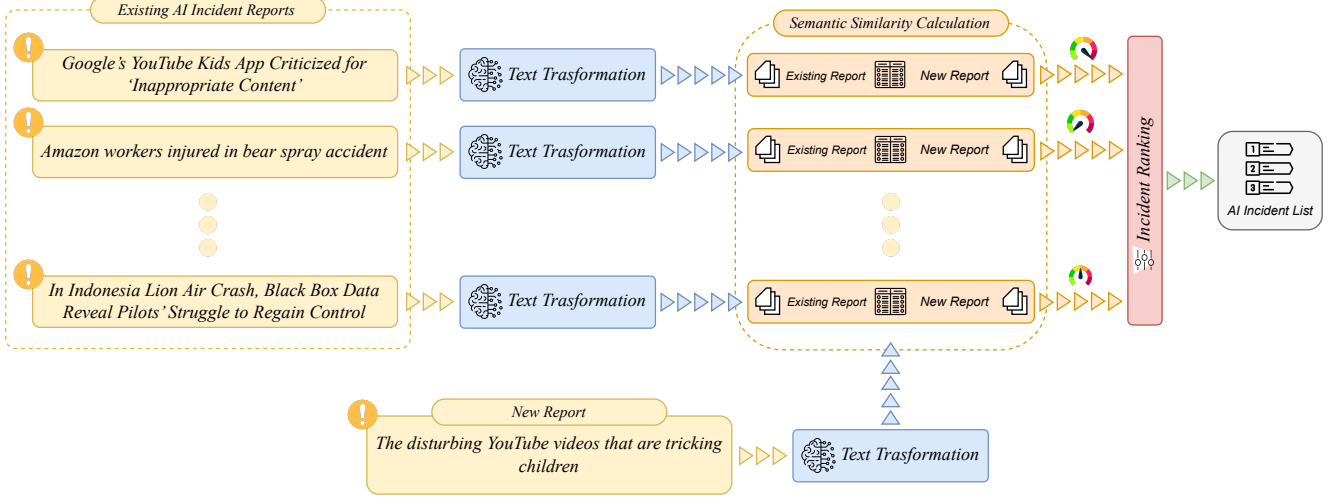
### 4.3 Impact of Title and Description Combination on Retrieval Performance (RQ1)

In the AIID system, reports are described by a title and an associated textual description. However, user-provided descriptions are only required to exceed a minimum length of 80 characters, a constraint that does not guarantee sufficient detail or informativeness for retrieval tasks. To address this limitation, we investigated the extent to which combining the title and description of each report enhances top- $K$  retrieval performance, particularly in scenarios where individual descriptions may lack sufficient detail.

As shown in Table 3, incorporating both the title and description consistently yields superior retrieval performance across all considered values of  $K$  and evaluation metrics. For instance, in terms of *Accuracy@K*, using only the title achieves a score of  $0.772 \pm 0.012$  at  $K = 3$ , while combining title and description raises the performance to  $0.982 \pm 0.006$ . A similar trend is observed at  $K = 5$ , where the title-only configuration attains an *Accuracy@K* of  $0.808 \pm 0.015$ , compared to  $0.987 \pm 0.005$  when both components are utilized.

Overall, these results emphasize that leveraging both title and description results in an average improvement across all evaluation criteria. Specifically, a 15% gain was observed in *Accuracy@K* with  $K = 10$ , while a peak improvement of 25% was observed in *MRR@K* with  $K = 3$ .

<sup>11</sup> The code will be made available upon acceptance.



**Figure 1. Overview of the semantic retrieval pipeline for linking new reports to existing AI Incidents.** Given a newly submitted report, the system ranks existing AI Incidents by computing semantic similarity between their textual representations. The pipeline consists of three main stages: (1) *Text Transformation*, where titles and descriptions are preprocessed and embedded; (2) *Semantic Similarity Calculation*, which computes pairwise similarity scores between the report and each incident; and (3) *Incident Ranking*, which generates a prioritized list of the most semantically relevant incidents.

K	Model	Accuracy@K	MRR@K	NDCG@K
3	BM25	0.584 ± 0.022	0.526 ± 0.019	0.526 ± 0.019
	BM25+T5	0.704 ± 0.022	0.637 ± 0.017	0.654 ± 0.017
	quora-roberta-base	0.513 ± 0.017	0.400 ± 0.012	0.429 ± 0.012
	ms-marco-MiniLM-L4-v2	0.666 ± 0.010	0.554 ± 0.009	0.583 ± 0.007
	multi-qa-distilbert-cos-v1	0.974 ± 0.018	0.948 ± 0.025	0.954 ± 0.023
	all-mpnet-base-v2	0.980 ± 0.004	0.955 ± 0.005	0.962 ± 0.004
	multi-qa-MiniLM-L6-cos-v1	<b>0.982 ± 0.006</b>	<b>0.963 ± 0.010</b>	<b>0.968 ± 0.008</b>
5	BM25	0.623 ± 0.019	0.535 ± 0.018	0.557 ± 0.018
	BM25+T5	0.745 ± 0.019	0.646 ± 0.016	0.671 ± 0.016
	quora-roberta-base	0.617 ± 0.018	0.424 ± 0.011	0.472 ± 0.012
	ms-marco-MiniLM-L4-v2	0.744 ± 0.007	0.572 ± 0.009	0.614 ± 0.008
	multi-qa-distilbert-cos-v1	0.982 ± 0.014	0.950 ± 0.023	0.958 ± 0.021
	all-mpnet-base-v2	0.985 ± 0.004	0.957 ± 0.004	0.965 ± 0.003
	multi-qa-MiniLM-L6-cos-v1	<b>0.987 ± 0.005</b>	<b>0.965 ± 0.010</b>	<b>0.970 ± 0.008</b>
10	BM25	0.666 ± 0.017	0.541 ± 0.017	0.571 ± 0.017
	BM25+T5	0.790 ± 0.014	0.653 ± 0.016	0.686 ± 0.015
	quora-roberta-base	0.741 ± 0.010	0.440 ± 0.011	0.505 ± 0.010
	ms-marco-MiniLM-L4-v2	0.837 ± 0.010	0.584 ± 0.009	0.645 ± 0.008
	multi-qa-distilbert-cos-v1	0.988 ± 0.009	0.950 ± 0.023	0.960 ± 0.019
	all-mpnet-base-v2	0.989 ± 0.002	0.957 ± 0.004	0.965 ± 0.003
	multi-qa-MiniLM-L6-cos-v1	<b>0.990 ± 0.005</b>	<b>0.965 ± 0.010</b>	<b>0.971 ± 0.008</b>

**Table 2. Performance comparison of baseline and embedding-based models across different values of K.** *multi-qa-MiniLM-L6-cos-v1* model consistently achieved the best performance across all values of K and evaluation metrics.

#### 4.4 Impact of Description Length on Retrieval Robustness (RQ2)

In the AIID system, reports are associated with textual descriptions, whose lengths can vary considerably. This variability introduces potential challenges, as excessively short descriptions may lack sufficient detail, while overly long descriptions may contain noise that negatively impacts retrieval performance. To address this, we examine if the length of descriptions influences performance, assessing

K	Input Type	Accuracy@K	MRR@K	NDCG@K
3	Incident Title	0.772 ± 0.012	0.707 ± 0.013	0.724 ± 0.013
	Incident Title + Description	<b>0.982 ± 0.006</b>	<b>0.963 ± 0.010</b>	<b>0.968 ± 0.008</b>
5	Incident Title	0.808 ± 0.015	0.715 ± 0.013	0.739 ± 0.014
	Incident Title + Description	<b>0.987 ± 0.005</b>	<b>0.965 ± 0.010</b>	<b>0.970 ± 0.008</b>
10	Incident Title	0.855 ± 0.021	0.722 ± 0.014	0.754 ± 0.015
	Incident Title + Description	<b>0.990 ± 0.005</b>	<b>0.965 ± 0.010</b>	<b>0.971 ± 0.008</b>

**Table 3. Retrieval performance comparison using title only versus title combined with description.** Combining title and description consistently improves retrieval effectiveness across all evaluation metrics and values of K, with gains ranging from 15% to 25% compared to using the title alone.

whether shorter or longer descriptions produce different results for our task.

Reports were stratified into subsets according to two distinct length-based partitioning criteria: one based on the median description length and the other on the 25th percentile. A report is classified as having a short description if its length falls below the respective threshold (median or 25th percentile), and as having a long description otherwise.

As shown in Table 4, shorter descriptions yield slightly better retrieval performance across all evaluation metrics. For instance, fixing  $K = 3$ , reports with descriptions shorter than the median achieve an  $MRR@K$  of  $0.979 \pm 0.010$ , compared to  $0.949 \pm 0.013$  for longer descriptions. A similar trend is observed when using the 25th percentile threshold: reports below this threshold reach an  $MRR@K$  of  $0.978 \pm 0.014$ , while those above attain  $0.959 \pm 0.013$ , respectively.

Although slight improvements are observed, the performance differences between shorter and longer descriptions remain minimal, suggesting that both types provide comparable retrieval effectiveness. This finding indicates that our framework maintains robust performance regardless of description length, demonstrating resilience to variability in the level of detail provided in the reports.

K	Length Condition	Accuracy@K	MRR@K	NDCG@K
3	< Median	<b>0.989 ± 0.004</b>	<b>0.979 ± 0.010</b>	<b>0.965 ± 0.003</b>
	≥ Median	0.973 ± 0.009	0.949 ± 0.013	0.955 ± 0.012
	< 25th Percentile	<b>0.988 ± 0.008</b>	<b>0.978 ± 0.014</b>	<b>0.981 ± 0.012</b>
	≥ 25th Percentile	0.979 ± 0.006	0.959 ± 0.010	0.964 ± 0.008
5	< Median	<b>0.992 ± 0.005</b>	<b>0.980 ± 0.010</b>	<b>0.983 ± 0.009</b>
	≥ Median	0.982 ± 0.008	0.951 ± 0.013	0.959 ± 0.011
	< 25th Percentile	<b>0.988 ± 0.008</b>	<b>0.978 ± 0.014</b>	<b>0.981 ± 0.012</b>
	≥ 25th Percentile	0.987 ± 0.005	0.961 ± 0.010	0.967 ± 0.008
10	< Median	<b>0.992 ± 0.005</b>	<b>0.980 ± 0.010</b>	<b>0.983 ± 0.009</b>
	≥ Median	0.987 ± 0.009	0.952 ± 0.013	0.960 ± 0.010
	< 25th Percentile	0.988 ± 0.008	<b>0.978 ± 0.014</b>	<b>0.981 ± 0.012</b>
	≥ 25th Percentile	<b>0.990 ± 0.006</b>	0.962 ± 0.010	0.969 ± 0.008

**Table 4. Retrieval performance across different description lengths.** Descriptions are classified as shorter or longer based on whether their length falls below or above the median or 25th percentile thresholds. Performance differences across these groups are minimal, suggesting that the retrieval framework is robust to input length variability.

#### 4.5 Impact of Training Data Size on Retrieval Performance (RQ3)

Until this point, our evaluation relied on static validation strategies with a fixed training set. However, considering that new reports are continuously reported over time, it becomes essential to assess how the model behaves as additional training data becomes available. To this end, we designed a progressive training protocol with fixed validation and test sets. Figure 2 illustrates the overall structure of the protocol. Specifically, the training set was first temporally ordered according to the reporting time of each instance. The ordered training set was then partitioned into five folds. Each fold  $F_i$  (for  $i = 1, \dots, 5$ ) includes both the instances from fold  $F_{i-1}$  and an additional batch of newly introduced instances. Formally, if each batch contains  $N$  new instances, the size of the  $i$ -th temporal fold is:

$$|F_i| = i \times N$$

with  $F_1$  consisting of  $N$  instances,  $F_2$  containing  $2N$  instances (the first  $N$  plus an additional  $N$ ), and so on. In our setup, each batch consists of 571 instances, resulting in fold sizes of 571, 1142, 1713, 2284, and 2855 instances, respectively.

We fine-tuned a separate model on each fold to evaluate how retrieval performance evolves as the size of the training set increases.

As shown in Table 5, the model demonstrates a consistent performance improvement as additional training data is incorporated, suggesting a beneficial effect from exposure to a growing number of incidents. For  $K = 3$ , accuracy increases from 0.921 in the first fold to 0.953 in the final one. Corresponding gains are observed in MRR, which rises from 0.859 to 0.926, and in NDCG, from 0.875 to 0.933. These trends indicate a progressively enhanced ranking capability as the training set expands over time. A similar pattern emerges for  $K = 5$ , where accuracy improves from 0.953 to 0.962 across folds. Similarly, MRR increases from 0.866 to 0.928, while NDCG advances from 0.888 to 0.937. Overall, these findings highlight the scalability and stability of our system under realistic data growth conditions.

K	Fold	Accuracy@K	MRR@K	NDCG@K
3	1	0.921	0.859	0.875
	2	0.940	0.878	0.894
	3	0.883	0.808	0.827
	4	0.953	0.925	0.932
	5	<b>0.953 (+3.47%)</b>	<b>0.926 (+7.80%)</b>	<b>0.933 (+6.63%)</b>
5	1	0.953	0.866	0.888
	2	0.968	0.885	0.906
	3	0.899	0.812	0.834
	4	0.972	0.929	0.940
	5	<b>0.962 (+0.95%)</b>	<b>0.928 (+7.16%)</b>	<b>0.937 (+5.52%)</b>
10	1	0.975	0.870	0.895
	2	0.990	0.888	0.913
	3	0.931	0.816	0.844
	4	0.987	0.931	0.944
	5	0.975 (+0.00%)	<b>0.930 (+6.90%)</b>	<b>0.941 (+5.14%)</b>

**Table 5. Retrieval performance across temporally expanding training folds.** Each fold incrementally introduces new training data in chronological order. Retrieval performance improves consistently across folds and metrics, highlighting the model’s stability to newly emerging incident reports.

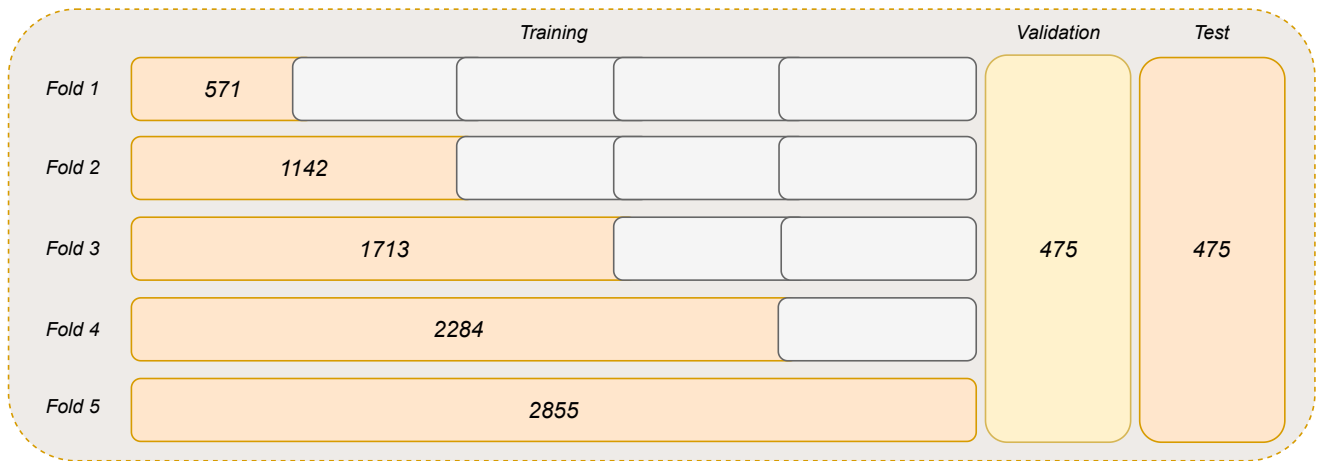
## 5 Conclusions, Limitations and Future Works

This work addresses the critical challenge of automating the association between newly reported reports and previously documented *AI Incidents* in the AI Incident Database. This task, currently performed manually, poses significant scalability limitations, potentially introducing also inconsistencies. To overcome these issues, we formalize the report-to-incident association as a semantic retrieval task and propose a framework that ranks historical incidents based on their semantic similarity to a given report.

We evaluate our approach using the AIID dataset, comprising 815 incidents and 3,805 reports. After benchmarking a range of retrieval models, we identify the *multi-qa-MiniLM-L6-cos-v1* sentence transformer as the most effective, consistently outperforming all baselines across Accuracy@K, MRR@K, and NDCG@K.

Our experimental analysis yields three key findings. First, combining the title and description of reports significantly improves retrieval performance over using the title alone, with gains ranging from 15% to 25% (RQ1). Second, the model maintains robust performance regardless of description length, with only marginal differences observed between shorter and longer inputs—highlighting the resilience of the framework to variability in input verbosity (RQ2). Third, our analysis shows that retrieval performance consistently improves as more training data becomes available, with gains observed across all evaluation metrics. This indicates that the proposed system scales effectively and remains robust as the incident repository grows over time (RQ3).

Despite these promising results, our framework relies exclusively on the textual content of reports, disregarding potentially valuable metadata such as incident categories, tags, or sources. This constraint may limit the model’s ability to capture deeper contextual or structural cues. Furthermore, by focusing solely on semantic similarity, the approach may overlook latent factors such as causal mechanisms, domain-specific nuances, or system-level attributes that are not explicitly described in the report text. Moreover, the current evaluation is restricted to pairwise comparisons between a single report and in-



**Figure 2. Evaluation protocol for assessing the impact of training data scale.** The training set is chronologically ordered and incrementally expanded in five cumulative folds, each adding a new batch of 571 newly reported reports. The validation and test sets remain fixed across all folds, enabling consistent evaluation of how retrieval performance changes as more data becomes available.

dividual candidate incidents, without considering the relational structure that may emerge from groups of reports referring to the same underlying incident.

Future work will explore the integration of structured metadata to complement text-based similarity, as well as methods for explainable retrieval to enhance transparency in the report-to-incident mapping. Another promising direction would be the integration of large language models (LLMs), leveraging their advanced reasoning and summarization capabilities to identify deeper semantic and causal links between reports that go beyond surface-level textual similarity. Finally, it would be valuable to evaluate our approach in the context of the actual AIID workflow, assessing its effectiveness in supporting editors during the real-time curation of incident reports.

## References

- [1] W. A. Addy, A. O. Ajayi-Nifise, B. G. Bello, S. T. Tula, O. Odeyemi, and T. Falaiye. Ai in credit scoring: A comprehensive review of models and predictive analytics. *Global Journal of Engineering and Technology Advances*, 18(02):118–129, 2024.
- [2] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy. “real attackers don’t compute gradients”: bridging the gap between adversarial ml research and practice. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML)*, pages 339–364. IEEE, 2023.
- [3] O. A. Bello, A. Ogundipe, D. Mohammed, F. Adebola, and O. A. Alonge. Ai-driven approaches for real-time fraud detection in us financial transactions: challenges and opportunities. *European Journal of Computer Science and Information Technology*, 11(6):84–102, 2023.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [5] E. Bondi-Kelly, T. Hartvigsen, L. M. Sanneman, S. Sankaranarayanan, Z. Harned, G. Wickerson, J. W. Gichoya, L. Oakden-Rayner, L. A. Celi, M. P. Lungren, et al. Taking off with ai: lessons from aviation for healthcare. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–14, 2023.
- [6] A. K. Chaurasia, C. J. Greatbatch, and A. W. Hewitt. Diagnostic accuracy of artificial intelligence in glaucoma screening and clinical practice. *Journal of Glaucoma*, 31(5):285–299, 2022.
- [7] H. Chopra, D. K. Shin, K. Munjal, K. Dhama, T. B. Emran, et al. Revolutionizing clinical trials: the role of ai in accelerating medical breakthroughs. *International Journal of Surgery*, 109(12):4211–4220, 2023.
- [8] G. Cohen. Algorithmic trading and financial forecasting using advanced artificial intelligence methodologies. *Mathematics*, 10(18):3302, 2022.
- [9] Z. Dai and J. Callan. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1533–1536, 2020.
- [10] G. Dartmann, A. Schmeink, V. Lücken, H. Song, M. Ziefle, and G. Prestifilippo. *Smart transportation: AI enabled mobility and autonomous driving*. CRC Press, 2021.
- [11] P. Diaconescu and V.-E. Neagoe. Credit scoring using deep learning driven by optimization algorithms. In *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE, 2020.
- [12] M. Feffer, N. Martelaro, and H. Heidari. The ai incident database as an educational tool to raise awareness of ai harms: A classroom exploration of efficacy, limitations, & future improvements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11, 2023.
- [13] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.
- [14] T. Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020.
- [15] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42, 2022.
- [16] M. Hurley and J. Adebayo. Credit scoring in the era of big data. *Yale JL & Tech.*, 18:148, 2016.
- [17] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.
- [18] S. E. Kase, C. P. Hung, T. Krayzman, J. Z. Hare, B. C. Rinderspacher, and S. M. Su. The future of collaborative human-artificial intelligence decision-making for mission planning. *Frontiers in Psychology*, 13: 850628, 2022.
- [19] O. Khattab and M. Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.
- [20] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.
- [21] M. Lu, C. Chen, and C. Eickhoff. Cross-encoder rediscovers a semantic variant of BM25. *CoRR*, abs/2502.04645, 2025. doi: 10.48550/ARXIV.2502.04645. URL <https://doi.org/10.48550/arXiv.2502.04645>.
- [22] G. Lupo. Risky artificial intelligence: The role of incidents in the path to ai regulation. *Law, Technology and Humans*, 5(1):133–152, 2023.
- [23] S. McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):15458–15463, May 2021. doi: 10.1609/aaai.v35i17.17817. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17817>.

- [24] S. McGregor, K. Paeth, and K. Lam. Indexing ai risks with incidents, issues, and variants, 2022. URL <https://arxiv.org/abs/2211.10384>.
- [25] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479, 2005.
- [26] P. Nie, Y. Zhang, X. Geng, A. Ramamurthy, L. Song, and D. Jiang. Decbert: Decoupling question and document for efficient contextual encoding. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1829–1832, 2020.
- [27] R. Nogueira, W. Yang, J. Lin, and K. Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019.
- [28] R. Nogueira, Z. Jiang, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.
- [29] R. Nogueira, Z. Jiang, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.
- [30] K. Paeth, D. Atherton, N. Pittaras, H. Frase, and S. McGregor. Lessons for editors of ai incidents from the ai incident database, 2024. URL <https://arxiv.org/abs/2409.16425>.
- [31] S. R. Pfohl, A. Foryciarz, and N. H. Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621, 2021.
- [32] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [33] N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.EMNLP-MAIN.365. URL <https://doi.org/10.18653/v1/2020.emnlp-main.365>.
- [34] C. Rigano. Using artificial intelligence to address criminal justice needs. *National Institute of Justice Journal*, 280(1-10):17, 2019.
- [35] S. E. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009. doi: 10.1561/15000000019. URL <https://doi.org/10.1561/15000000019>.
- [36] R. Rodrigues, A. Resseguier, and N. Santiago. When artificial intelligence fails: The emerging role of incident databases. *Pub. Governance, Admin. & Fin. L. Rev.*, 8:17, 2023.
- [37] G. Rosa, L. H. Bonifacio, V. Jeronymo, H. Q. Abonizio, M. Fadaee, R. A. Lotufo, and R. Nogueira. In defense of cross-encoders for zero-shot retrieval. *CoRR*, abs/2212.06121, 2022. doi: 10.48550/ARXIV.2212.06121. URL <https://doi.org/10.48550/arXiv.2212.06121>.
- [38] K. J. Ruskin, C. Corvin, S. Rice, G. Richards, S. R. Winter, and A. Clebone Ruskin. Alarms, alerts, and warnings in air traffic control: An analysis of reports from the aviation safety reporting system. *Transportation Research Interdisciplinary Perspectives*, 12:100502, 2021. ISSN 2590-1982. doi: <https://doi.org/10.1016/j.trip.2021.100502>. URL <https://www.sciencedirect.com/science/article/pii/S2590198221002074>.
- [39] J. D. Schloetzer and K. Yoshinaga. Algorithmic hiring systems: Implications and recommendations for organisations and policymakers. In *YSEC Yearbook of Socio-Economic Constitutions 2023: Law and the Governance of Artificial Intelligence*, pages 213–246. Springer, 2023.
- [40] K. Song, X. Tan, T. Qin, J. Lu, and T. Liu. Mpnnet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html>.
- [41] A. Talmor and J. Berant. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. *arXiv preprint arXiv:1905.13453*, 2019.
- [42] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [43] X. Zhao, J. Liu, Y. Zhang, Z. Yu, and B. Guo. Haiformer: Human-ai collaboration framework for disease diagnosis via doctor-enhanced transformer. In *ECAI 2024*, pages 1495–1502. IOS Press, 2024.