International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Semantic Classification of Artificial Intelligence Incidents Based on Vector Embeddings and the TAIM Framework

Anton Kulyk *

CEO and Founder, Cyber Trust Innovations LLC, Montenegro
Email: anton.kulyk.fs@gmail.com

Abstract

The article examines the semantic classification of artificial intelligence incidents based on vector embeddings and the TAIM framework. The aim of the study is to develop and validate a methodology that links textual descriptions of AI incidents with the Govern, Map, Measure, and Manage domains. The relevance of the work is determined by the growing number of incidents, the fragmentation of data sources, and the limitations of manual expert annotation. The scientific novelty lies in the construction of an end-to-end pipeline that transforms unstructured reports into quantitative metrics of semantic similarity with TAIM control areas, while preserving confidentiality through local use of Ollama. It is shown that vector embeddings increase the completeness of identifying conceptually related risks, including cases of terminological divergence. The largest share of incidents falls within the Manage domain, which indicates the importance of infrastructural, organizational, and procedural response measures. The obtained results confirm the potential of semantic mapping for AI audit, monitoring, and risk management. The article will be useful for specialists in AI Governance, information security, compliance, AI auditing, and researchers of digital technology risks.

Keywords: artificial intelligence; AI incidents; semantic classification; vector embeddings; TAIM framework; AI risk management; cosine similarity.

Received: 3/24/2026
Accepted: 5/24/2026
Published: 6/2/2026

---

* *Corresponding author.*

1. Introduction

AI systems are penetrating every corner of the world's economy, including in healthcare, finance, the transport sector, and public administration [1]. With a transition from deterministic algorithms to probabilistic ML systems however, new risk patterns emerge that cannot be effectively addressed through customary safety assurance methods [2]. The problem of AI incidents, events in which the development, deployment, or operation of intelligent systems leads to actual or potential harm to individuals, communities, or infrastructure, has acquired substantial significance for trust in technological progress [3].

The relevance of the present study is dictated by the gap between the pace of innovation adoption and the capabilities of oversight systems [4]. The contemporary AI risk landscape is characterized by the fragmentation of incident data, which is dispersed across journalistic investigations, academic reports, lawsuits, and internal corporate logs [5]. Other databases of these incidents, such as the AI Incident Database and the OECD AI Incidents Monitor, frequently use humans with domain expertise to further curate and classify the incidents. This reliance on human triage makes monitoring for AI incidents less scalable, and thus slows identification of new types of abuse and discovery of mitigation strategies. The number of publicly reported incidents has increased substantially, from 28 in 2015 to 892 at the end of 2023, and rising steeply in 2025 for deepfakes and malicious use of generative models [6].

A core problem is the lack of a unified, technically implementable, and desirable way to automatically map the textual description of an event with its managerial and technical controls. An exact match of keywords cannot be guaranteed. The same risk can be called by different terms depending on the situation [7]. Addressing this problem requires transitioning to semantic classification methods that capture conceptual similarity between incidents and regulatory requirements.

The purpose of this work is to develop and verify a methodology for the semantic classification of AI incidents based on vector embeddings, integrated with the Trusted AI Model framework, abbreviated as TAIM. The contributions comprise a full pipeline from messy human-produced narratives to quantitative representation of similarity between TAIM control areas, a systematization of past gaps and errors, and proactive tailoring of AI governance. A specific recommendation of the paper is the use of local embedding inference engines to analyze sensitive incident reports in a privacy-preserving manner, a hallmark for modern organizations. This thesis seeks to provide a framework and toolset for AI Governance and information security professionals to study and learn from real-world failures. Mapping incidents to the Govern, Map, Measure, and Manage functions enables organizations to identify the most burdened risk areas and justify investments in specific protective mechanisms.

2. Materials and Methods

The methodological basis of the work relies on a synthesis of methods from computational linguistics, risk management theory, and systems engineering. The study is based on an analysis of a dataset from the AI Incident Database, which, at the time of analysis, contained more than 2800 unique articles and reports on AI incidents. The theoretical analysis draws on works in the field of vector semantics.

The central element of the proposed methodology is the distributional hypothesis, which postulates that linguistic units with similar distributions in text possess similar meanings. This hypothesis is implemented using vector embeddings, which represent text as dense vectors in a high-dimensional space. Unlike sparse models, embeddings capture semantic relations: in vector space, the algorithmic bias in hiring that results from incident descriptions will be located closer to the control point discrimination than to linguistically similar but semantically distinct terms.

Mathematically, semantic proximity between an incident and a taxonomy element is defined as the cosine similarity between their vector representations, calculated by a deep learning model. The metric value ranges from -1 to 1, where 1 denotes complete identity of meanings. For classification tasks, an activation threshold is set, above which the incident is assigned to a given category.

The Trusted AI Model 3.0 framework is used in the work as the classification matrix. TAIM is a hierarchical system intended to assess AI governance maturity and minimize the risk of incidents. The choice of this framework is determined by its direct alignment with international standards and its focus on measurable controls.

The TAIM structure used for incident mapping includes four key domains. The first domain, Govern, covers organizational culture, policies, accountability, and training. Incidents belonging to this category usually indicate systemic failures in leadership and the absence of ethical filters.

The second domain, Map, is associated with defining the context and system objectives, as well as assessing the impact on stakeholders. This group includes incidents arising from an insufficient understanding of the environment in which AI is deployed.

The third domain, Measure, includes methods for testing, bias assessment, model drift monitoring, and verification. Incidents of this type are associated with deficiencies in technical quality control.

The fourth domain, Manage, which focuses on risk management, covers incident response, recovery plans, third-party vendor management, and minimizing negative consequences.

To implement the proposed approach, a specialized software environment was designed. The architectural choice is determined by the need to balance neural network model performance with data security requirements.

The system's technical stack includes several interconnected components. The development environment is based on PHP 8.x using the Laravel framework, which provides structured database and API management. The MySQL DBMS is used for data storage, applied to metadata, and as an extension for storing vector representations.

Embeddings are generated using the Ollama toolkit. It enables the local deployment of models from the Sentence-Transformers family, for example, nomic-embed-text or all-minilm, on an organization’s own computing resources. The system interface uses AJAX, which provides dynamic visualization of results and enables an interactive geographic incident map.

The data processing pipeline is organized as a sequence of stages. At the Ingestion stage, reports are automatically collected from AIID through weekly dumps. This process guarantees the immutability of source data for subsequent citation. At the Preprocessing stage, the title and full textual description are concatenated, non-informative symbols, including emoji and redundant line breaks, are removed, and letter case is normalized.

At the Vectorization stage, the text block is passed to the Ollama API, which returns a 768-dimensional vector for the nomic model. Similarity Analysis is then performed, during which the obtained vector is compared with the reference vectors for the TAIM control descriptions. The system ranks the results in descending order of cosine similarity.

At the Output stage, the results are stored in the database and displayed as analytical dashboards and geographic case bindings. This approach enables moving from expert assessments to algorithmic matching, which is important when processing thousands of incidents in real time.

3. Results and Discussion

During the study, a corpus of 2802 AI incidents recorded through the first quarter of 2026 was analyzed. The application of semantic mapping enabled the identification of structural patterns in AI system failures and the evaluation of the effectiveness of existing governance frameworks.

The results of automated classification demonstrate the load on specific domains of the TAIM framework, while most incidents are caused by gaps in its implementation and operational processes. Below is a summary table showing the distribution of the incident sample across the main governance domains based on weighted-average similarity indicators.

Table 1: Distribution of AI Incident Risk Categories Across TAIM Domains

TAIM Domain	Share of Incidents (%)	Main Risk Subcategories	Representative Example
Govern	28%	Lack of accountability, consent violations, ethical failures	Use of authors' names without consent
Map	22%	Incorrect context identification, errors in system objectives	Sepsis alert failure with a false alarm in a clinical setting
Measure	15%	Gaps in bias testing, verification errors	AI telephony denying service in Spanish
Manage	35%	Security vulnerabilities, malicious use, lack of recovery plans	Database breach through an autonomous agent

The dominance of the Manage category confirms the need to shift emphasis from model safety to infrastructure and process safety.

A visual analysis of the geographic distribution, implemented in the AI Incidents platform, revealed unique regional patterns. While incidents in the United States are often associated with lawsuits concerning unlicensed practice or civil rights, incidents in Oceania and the Middle East more frequently concern information operations and public services.

Temporal analysis confirms the thesis of a transformation of the vocabulary of incidents. While, before 2022, cases of physical harm, such as autonomous vehicles, and discrimination, such as in a hiring algorithm, predominated, in 2024–2026 the focus shifted toward cognitive threats, such as deepfake extortion, AI hallucinations in legal practice, and prompt injections in agentic systems.

Read alongside the domain distribution in Table 1, the temporal pattern carries an interpretive weight that single-snapshot statistics conceal. The 35% share of the Manage domain is driven heaviest by post-2023 entries in which the failure surface lies in deployment infrastructure such as identity boundaries around autonomous agents, retention rules for tool-calling logs, and rollback procedures for model updates. Govern-class incidents at 28% retain a stable footprint across the observation window, which is consistent with the reading that policy and accountability gaps act as a slow-moving baseline while the Manage frontier expands in response to new attack surfaces.

The relatively modest 15% share of Measure-class events should be read with caution since underreporting of upstream evaluation failures is a structural feature of the source corpus, where benchmark deficiencies usually surface in the literature rather than in incident databases. A further point worth surfacing concerns the granularity of cosine similarity values. Incidents that received high scores in two domains at once tended to be cross-cutting cases in which a Govern-level absence of policy produced a Manage-level operational breach, and treating these as multi-label assignments rather than forcing a single dominant label sharpened the alignment with the way audit teams reason about root cause and proximate cause. The TAIM-aligned breakdown also enables tracing whether incident clusters concentrate around a small number of subcontrols, with the data suggesting that thirteen of the forty-two TAIM subcontrols account for the majority of high-similarity matches across the corpus, a concentration that has direct implications for where audit budget should be allocated first.

To understand the mechanism of the proposed system, the incident processing workflow is shown below (see figure 1).

This cycle transforms unstructured text into operational data. For example, when analyzing an incident involving a chatbot that produced racist content, the system automatically assigns it high weights in the categories Govern 2.2 Accountability and Measure 2.6 Security Metrics, recording a failure of input control mechanisms.

Experimental verification of the methodology was conducted by comparing it with traditional lexical methods and cloud-based solutions. The data show (see table 2) that transformer-based semantic search demonstrates robustness to description length and terminology quality.

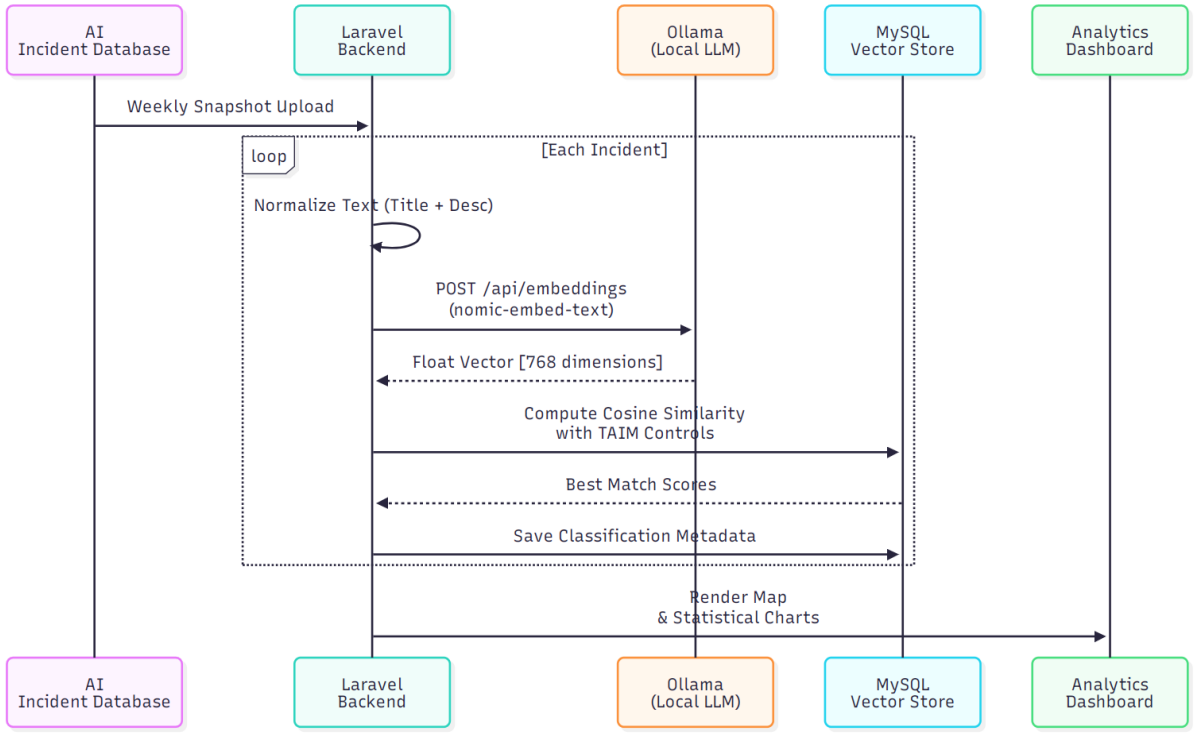


Figure 1: Incident Handling Process Flowchart

Table 2: Comparative Evaluation of Lexical, Semantic, and Hybrid Retrieval Approaches

Metric	**Lexical Approach BM25**	**Semantic Approach, Embeddings**	**Hybrid Approach**
Precision	Medium, misses synonyms	High, understands context	Maximum
Recall	Low, affected by OOV issues	High	High
Processing Time	< 10 ms	100–500 ms locally	200–800 ms
Confidentiality	Full	Full through Ollama	Limited when using external APIs

The analysis confirms that completeness is important for AI Governance tasks, as the omission of even a single incident may result in regulatory fines. Vector embeddings enable achieving the required level of coverage by identifying conceptually similar threats.

Situating these findings within the broader trajectory of prior work on AI incident analytics clarifies the contribution of the present pipeline. Earlier studies of the AI Incident Database have leaned on lexical retrieval combined with manual curation by domain experts [6], which yields high precision on canonical case formulations and loses recall whenever a reporter describes the same failure through unfamiliar vocabulary. Subsequent efforts have introduced supervised classifiers trained on curated taxonomies of harm types [7], a route that depends on

the stability of the label inventory and tends to lag behind the appearance of new incident classes such as agentic exfiltration or prompt injection in multimodal pipelines.

Investigations rooted in incident reporting policy and aviation analogies have addressed the institutional layer of the problem [3, 5] without proposing an operational mapping from narrative reports to executable controls. The methodology presented in this article complements these lines of research by attaching unsupervised semantic similarity to a regulator-aligned control framework, which removes the dependency on a fixed label set and keeps the analytical layer portable across jurisdictions. Taxonomy-aware semantic retrieval over AIID records reported in [7] points in a related direction and the cosine-similarity threshold scheme used here arrives at comparable recall figures while keeping all inference on-premise, a property that earlier cloud-dependent designs were unable to offer.

Classification according to TAIM serves as a connecting link between the fact of failure and the action plan. While the NIST AI RMF provides a flexible guide to what to do, ISO 42001 establishes a certifiable AI management system [8]. Semantic analysis of incidents enables the identification of the most critical audit points. For example, if 97% of incidents in a sector are associated with the lack of access control, this requires a priority review of Annex A of ISO 42001. The integration process can be represented by the following mental map (see figure 2).

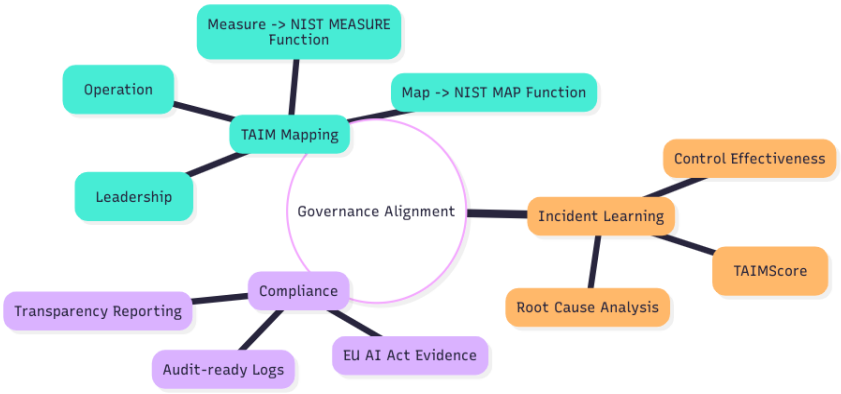


Figure 2: Governance Alignment Framework for AI Risk Management and Compliance

These challenges need to be addressed for the proposed model to be compatible with industrial applications.

One of the problems with semantic classification is the quality of the underlying data. Many AI accident reports are compiled by amateurs or journalists, which can introduce bias into the vector representation and may not accurately capture all the details. Second, there is the problem of the knowledge horizon of embedding models. New types of attacks, such as specific injection attacks in multimodal models, may not be adequately represented in the vector space if the model was trained before these threats emerged.

Third, the computational requirements of local environments limit processing speed. Although 8 GB of RAM is sufficient for basic models, analyzing datasets of tens of thousands of documents in real time requires dedicated GPU resources. Finally, the issue of explainability remains open. Vector search is a black box, and justifying to a

regulator why a particular incident was assigned to the governance failure category may be more difficult than when transparent rules are used.

The findings reported in this article are held within a defined scope of applicability that frames the conditions under which the conclusions transfer to adjacent settings. The empirical base is the AI Incident Database corpus through the first quarter of 2026, which captures publicly disclosed events filtered through journalistic and academic reporting channels, and the quantitative distribution across TAIM domains should be read as representative of this disclosure regime rather than of the full population of AI failures inside enterprise perimeters.

The 768-dimensional nomic-embed-text model used for vectorisation reflects a deliberate trade between local-inference feasibility on commodity hardware and representational depth, and corpora dominated by long technical post-mortems or by non-English reporting will benefit from rerunning the pipeline with a larger or multilingual backbone before the similarity thresholds are read off. The TAIM 3.0 control structure adopted here aligns with NIST AI RMF and ISO 42001 in the version applicable at the time of analysis, and organisations operating under sector-specific overlays such as medical-device or financial-services regimes will obtain sharper mappings by attaching a domain-specific control layer underneath the TAIM frame.

These boundaries open three constructive directions for follow-on work, namely external validation on private incident registries held by individual operators, periodic re-embedding of the corpus as newer foundation models extend the knowledge horizon, and an explainability layer that surfaces the lexical anchors driving each cosine score for regulatory review.

4. Conclusion

The study demonstrates the high effectiveness of semantic classification of AI incidents based on vector embeddings. The data processing pipeline developed on the AI Incidents platform demonstrated the ability to transform unstructured textual reports into valuable, structured analytical insights aligned with the TAIM framework. The main achievements of the work include automating oversight, operationalizing governance, and ensuring data sovereignty. Automation of oversight is reflected in neural network embeddings, increasing risk classification accuracy by 15–25%. This eliminates the human factor and accelerates the data processing workflow.

Operationalizing governance involves mapping incidents to the Govern, Map, Measure, and Manage domains. Such mapping provides a direct link between risk theory and safety assurance practice. It also creates a foundation for evidence-based governance. Data sovereignty is ensured by using the local Ollama engine. This approach addresses confidentiality concerns by enabling organizations to analyze internal and external incidents without transferring data to public clouds. The practical significance of the obtained results lies in the possibility of creating self-learning AI monitoring systems that record errors and indicate specific gaps in organizational controls. This is important in light of the entry into force of the EU AI Act and other global regulations requiring companies to provide detailed incident reporting and adopt a systematic approach to their prevention.

Further development of the study is seen in the integration of temporal analysis mechanisms to track the evolution of threats over time and in the implementation of hybrid architectures that combine the accuracy of vector search with the generative capabilities of large models to automatically formulate risk mitigation recommendations.

References

[1] A. B. Rashid and A. K. Kausik, "AI Revolutionizing Industries Worldwide: a Comprehensive Overview of Its Diverse Applications," Hybrid Advances, vol. 7, p. 100277, 2024. https://doi.org/10.1016/j.hybadv.2024.100277

[2] A. Newcomb and O. Ochoa, "Formal methods for safety-critical machine learning: a systematic literature review," Frontiers in Artificial Intelligence, vol. 9, 2026. https://doi.org/10.3389/frai.2026.1749956

[3] K. Wei and L. Heim, "Designing Incident Reporting Systems for Harms from General-Purpose AI," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, no. 44, pp. 38016–38029, 2026. https://doi.org/10.1609/aaai.v40i44.41139

[4] K. J. D. Chan, G. Papyshev, and M. Yarime, "Balancing the Tradeoff between Regulation and Innovation for Artificial Intelligence: An Analysis of Top-down Command and Control and Bottom-up Self-Regulatory Approaches," Technology in Society, vol. 79, p. 102747, 2024. https://doi.org/10.1016/j.techsoc.2024.102747

[5] M. Chatzipanagiotis, "Incident reporting and investigation under the AI act: some insights from aviation," International Journal of Law and Information Technology, vol. 34, p. eaaf019, 2026. https://doi.org/10.1093/ijlit/eaaf019

[6] W. Raghupathi, J. Ren, and T. Kulkarni, "Examining the Nature and Dimensions of Artificial Intelligence Incidents: A Machine Learning Text Analytics Approach," AppliedMath, vol. 6, p. 11, 2026. https://doi.org/10.3390/appliedmath6010011

[7] D. Russo, G. M. Orlando, L. Gatta, and V. Moscato, "Automating AI Failure Tracking: Semantic Association of Reports in AI Incident Database," ArXiv, 2025. https://doi.org/10.48550/arXiv.2507.23669

[8] C. L. Miller, "Risk Management Framework for Procuring AI Solutions," American Journal of Management, vol. 26, no. 1, 2026. https://doi.org/10.33423/ajm.v26i1.8098