Cybersecurity and Cybercrime 2026/1[10] DOI: 10.5604/01.3001.0055.8282

OPERATIONALISING CYBER RISK MANAGEMENT: CONNECTING CYBER INCIDENTS TO MITRE ATT&CK TECHNIQUES, CONTROLS, and METRICS

M.Sc. Emad SHERIF[1]
Researcher at De Montfort University, Leicester, UK
Dr Iryna YEVSEYEVA[2]
Associate Professor at De Montfort University, Leicester, UK
Prof Vitor BASTO-FERNANDES[3]
Professor at University Institute of Lisbon, ISTAR-IUL, Lisboa, Portugal
Prof Allan COOK[4]
Professor at De Montfort University, Leicester, UK

Summary
The escalating frequency of cyber-attacks poses significant challenges for organisations, particularly small enterprises constrained by limited financial resources. This research presents a novel framework leveraging a knowledge base and Natural Language Processing to address these challenges through automated mapping of cyber incidents to adversary techniques. We introduce the Cyber Catalog, a comprehensive knowledge base that systematically integrates CIS Critical Security Controls, MITRE ATT&CK techniques, and quantifiable metrics, enabling organisations to connect threat intelligence directly to actionable controls and measurable outcomes. To operationalise the framework, we fine-tuned a general-purpose sentence transformer on an augmented dataset of 74,986 incident-technique pairs to enhance semantic similarity between cyber incidents and MITRE ATT&CK techniques. Our fine-tuned model achieved a Spearman correlation of 0.7894 and Pearson correlation of 0.8756, representing improvements of approximately 0.21 over baseline models including all-mpnet-base-v2, all-distilroberta-v1, and all-MiniLM-L12-v2. The model also exhibited significantly lower prediction errors (MAE = 0.135, MSE = 0.027), confirming superior accuracy and consistency with three-fold reductions compared to baselines. This work bridges the gap between threat intelligence and operational security management, providing an actionable tool for cyber incident response and evidence-based risk management.
Keywords: CIS Security Controls, Cyber Incidents, MITRE ATT&CK, Natural Language Processing, Metrics.

1 ORCID 0000-0002-8450-338X
2 ORCID 0000-0002-1627-7624
3 ORCID 0000-0003-4269-5114
4 ORCID 0000-0002-0041-3724

1. INTRODUCTION

The contemporary cyber threat landscape is characterised by an unprecedented escalation in both the frequency and sophistication of cyber-attacks. Organisations of all sizes face threats from advanced persistent threat (APT) groups and opportunistic attackers exploiting vulnerabilities. However, the burden of these threats falls disproportionately on small and medium-sized enterprises (SMEs), which often lack the financial resources, technical expertise, and personnel necessary to implement robust cyber security programmes.

Traditional approaches to cyber incident analysis are manual and demand substantial time investment from security analysts. These processes involve examining unstructured incident descriptions, identifying the adversary techniques, mapping these techniques to relevant security controls, and assessing the effectiveness of implemented controls. This manual process is not only time-consuming but also prone to error and inconsistency, particularly when dealing with the volume and variety of incidents that organisations face.

The MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) framework has emerged as the de facto standard for describing adversary behaviour in the cyber security domain[5]. It provides a comprehensive, structured taxonomy of adversary techniques based on real-world observations. Similarly, the Centre for Internet Security (CIS) Critical Security Controls offer a prioritised set of actions that collectively form a defence-in-depth security posture[6]. However, the manual process of connecting cyber incidents to ATT&CK techniques and subsequently to CIS controls remains a significant operational bottleneck.

Furthermore, whilst organisations implement various security controls, monitoring and assessing such controls often relies on subjective judgement rather than objective, quantifiable metrics. The absence of measurable indicators makes it difficult for organisations to determine whether their security measures are yielding the desired risk reduction[7]. Research has demonstrated that effective security metrics must adhere to SMART criteria (Specific, Measurable, Achievable, Relevant, and Time-bound) to provide actionable insights for decision-makers. Without such metrics, organisations struggle to justify security expenditures, prioritise improvements, or demonstrate compliance[8].

Current cyber incident analysis workflows face several interconnected challenges. Translating unstructured incident descriptions into structured ATT&CK technique classifications requires considerable time investment, and once techniques are identified,

5 MITRE. (2023). MITRE ATT&CK. https://attack.mitre.org, Accessed 2 June 2026.
6 Centre for Internet Security. (2021). CIS Critical Security Controls, https://www.cisecurity.org/controls/v8, Accessed 2 June 2026.
7 Savola, R. M. (2013). Quality of security metrics and measurements. Computers and Security, 37, 78–90.
8 Hubbard, D. W., & Seiersen, R. (2023). How to measure anything in cybersecurity risk (2nd ed.). Wiley.

analysts must manually determine which security controls are relevant, a process that demands familiarity with both the ATT&CK framework and control frameworks such as CIS Controls. Compounding this, assessing and monitoring implemented controls typically relies on subjective judgements rather than objective, quantifiable metrics, making it difficult for organisations to determine whether their security measures are yielding the desired risk reduction. Furthermore, manual processes cannot efficiently manage the scale of modern cyber incidents, resulting in delayed response times and potential gaps in security posture. Collectively, these challenges stress the need for automated, scalable approaches that can facilitate incident-to-control mapping whilst accurately matching relevant techniques.

Hence, this work addresses the following research questions:

- RQ1: What are the methodological challenges in constructing comprehensive mappings between CIS security controls and MITRE ATT&CK techniques, and what objective metrics can be defined to monitor the controls?
- RQ2: To what extent can a general-purpose transformer model accurately map cyber incidents to MITRE ATT&CK techniques, and what improvements are achievable through fine-tuning?
- RQ3: How can automated mappings between cyber incidents and security controls be leveraged to support risk-based decision-making?

Thus, the primary objectives of this work are threefold. First, to accelerate incident triage through automated identification of relevant adversary techniques using a globally accessible knowledge base. Second, to improve cyber risk management by enabling automated incident-to-control linkage. Third, to assess and monitor control implementation using quantifiable metrics.

We hypothesise that developing a framework which automatically maps cyber incidents to MITRE ATT&CK techniques at scale, integrated with relevant security controls that are assessed objectively through quantifiable metrics, will provide decision-makers with actionable intelligence for data-informed cyber risk mitigation. Such a framework should reduce organisational risk exposure, enhance cyber resilience, and strengthen overall security posture.

Therefore, we introduce the Cyber Catalog, a comprehensive knowledge base that systematically integrates three critical cyber security components: Security Controls, ATT&CK techniques, and metrics. The principal contribution of this work is application-oriented and integrative: rather than proposing a fundamentally new machine-learning algorithm, the study combines domain-specific semantic mapping, ATT&CK-to-control linkage, and metrics-driven risk assessment into a unified workflow for operational cyber risk management. The Cyber Catalog addresses a fundamental gap in operational cyber security by providing practitioners with a unified reference that connects security measures to threat intelligence and measurable outcomes.

We also present a comprehensive approach for creating high-quality training data using large language models (LLMs) with rigorous quality assessment based on

BERTScore evaluation[9]. We address the challenge of one-to-many mappings in training data through a modified loss function that prevents false negative penalties during training, combined with hard negative mining to enhance discrimination[10].

Furthermore, we demonstrate substantial improvements achieved through fine-tuning a general-purpose sentence transformer on domain-specific cyber incident data, with improvements of approximately 0.21 in Spearman correlation over baseline models. We also provide extensive comparative analysis against top baseline models using both correlation-based and error-based metrics, demonstrating three-fold reductions in prediction errors.

The remainder of this paper is organised as follows: Section 2 reviews related work. Section 3 details our methodology, including data preparation. Section 4 presents our experimental results. Section 5 discusses the implications, limitations, and potential applications of our approach. Section 6 concludes the paper and outlines directions for future research.

2. Related Work

2.1 Cyber Risk Management and Threat Intelligence

Cyber risk management has evolved significantly over the past decade, transitioning from reactive, signature-based defence mechanisms to proactive, intelligence-driven approaches. Contemporary frameworks emphasise continuous monitoring, predictive analytics, and risk-based prioritisation. The integration of threat intelligence with controls represents a critical capability for organisations seeking to move beyond purely defensive postures.

The MITRE ATT&CK framework has become the cornerstone of modern threat intelligence, providing a common language for describing adversary behaviour. Efforts have explored methods for automatically extracting ATT&CK techniques from unstructured text sources, including threat reports. However, most existing approaches rely on rule-based systems or traditional machine learning classifiers that struggle with the semantic nuances and contextual variations present in real-world incident data[11,12].

9 Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1904.09675.
10 Robinson, J., Chuang, C.-Y., Sra, S., & Jegelka, S. (2021). Contrastive learning with hard negative samples. Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).

11 Mohasseb, A., Aziz, B., Jung, J., & Lee, J. (2019). Predicting Cyber Security Incidents Using Machine Learning Algorithms: A Case Study of Korean SMEs. In Proceedings of the 5th International Conference on Information Systems Security and Privacy, 230-237.
12 Rafiey, P. & Namadchian, A. (2025). Mapping Vulnerability Description to MITRE ATT&CK Framework by LLM. Advances in Artificial Intelligence and Machine Learning. 2025;5(3):243.

2.2 Natural Language Processing

The application of NLP (Natural Language Processing) techniques to cyber security tasks has gained considerable attention in recent years. Early work focused on named entity recognition for extracting indicators of compromise from text[13]. More recent research has explored document classification, relationship extraction, and semantic similarity tasks specific to the cyber security domain[12].

Transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers) and its variants, have demonstrated remarkable performance across a wide range of NLP tasks[14]. Sentence-BERT and similar sentence embedding models extend these capabilities to semantic similarity tasks by producing fixed-length vector representations that capture sentence-level semantics. These models have been successfully applied to tasks such as semantic search, duplicate question detection, and document clustering[14,15,16].

2.3 Fine-Tuning

The practice of adapting pre-trained models to specific domains or tasks, has become the dominant paradigm. General-purpose language models trained on large corpora can be fine-tuned on domain-specific data to achieve substantial performance improvements over baseline models[17,18]. This approach is particularly valuable in specialised domains such as cyber security, where labelled training data may be scarce or expensive to obtain.

Several studies have investigated the fine-tuning of transformer models for cyber security tasks[19,20,21]. However, most existing work focuses on classification tasks rather

13 Georgescu, T. M., Iancu, B., & Zurini, M. (2019). Named-entity-recognition-based automated system for diagnosing cybersecurity situations in IoT networks. Sensors (Switzerland), 19(15). https://doi.org/10.3390/s19153380.

14 Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982–3992. https://doi.org/10.18653/v1/D19-1410.

15 Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 1–14.

16 Pavlyshenko, B., & Stasiuk, M. (2025). Semantic similarity analysis using transformer-based sentence embeddings. Electronics and Information Technologies, 30. https://doi.org/10.30970/eli.30.4.

17 Stankevičius, L., & Lukoševičius, M. (2024). Extracting Sentence Embeddings from Pretrained Transformer Models. Applied Sciences (Switzerland), 14(19). https://doi.org/10.3390/app14198887.

18 Sufi, F. (2024). Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. Information (Switzerland), 15(2). https://doi.org/10.3390/info15020099.

19 Bayer, M., Frey, T., & Reuter, C. (2023). Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. Computers and Security, 134. https://doi.org/10.1016/j.cose.2023.103430.

20 Pavlyshenko, B., & Stasiuk, M. (2025). Semantic similarity analysis using transformer-based sentence embeddings. Electronics and Information Technologies, 30. https://doi.org/10.30970/eli.30.4.

21 Stankevičius, L., & Lukoševičius, M. (2024). Extracting Sentence Embeddings from Pretrained Transformer Models. Applied Sciences (Switzerland), 14(19). https://doi.org/10.3390/app14198887.

than semantic similarity. Furthermore, limited research has addressed the specific challenge of mapping unstructured incident descriptions to structured technique taxonomies such as MITRE ATT&CK.

2.4 LLM-Based ATT&CK Mapping

Recent work has increasingly explored the use of large language models (LLMs) for cyber threat intelligence tasks, including the mapping of vulnerability descriptions, threat reports and incident narratives to MITRE ATT&CK techniques. For example, Rafiey and Namadchian investigated the use of general-purpose LLMs, including GPT-based models, for mapping CVE descriptions to MITRE ATT&CK techniques, highlighting the potential of prompt-based approaches to reduce the need for repeated model retraining as vulnerabilities and ATT&CK versions evolve[22]. Similarly, recent work on extracting ATT&CK techniques from cyber threat intelligence reports has examined LLM-based and LLM-assisted pipelines, including approaches that combine LLM summarisation, rebalanced datasets and retrained domain models to improve technique identification[23].

LLM-based cyber threat intelligence systems have also been explored in broader attack-level intelligence extraction. IntelEX, for instance, uses LLMs together with an external intelligence vector database to extract structured attack-level intelligence from threat reports, including tactics, techniques and procedures, while also incorporating an LLM-based judgement module to reduce hallucination and false positives[24]. In addition, the MITRE Center for Threat-Informed Defense's Threat Report ATT&CK Mapper (TRAM) project demonstrates the continuing practical relevance of automated ATT&CK mapping, with recent versions making use of fine-tuned LLMs to improve the speed and accuracy of TTP identification in cyber threat intelligence reports[25].

Despite these advances, much of the existing work treats ATT&CK mapping as the main endpoint of the analysis. Less attention has been given to connecting identified adversary techniques to defensive controls and measurable indicators. This creates a gap between threat-intelligence interpretation and operational security management. The present work addresses this gap by positioning incident-to-technique mapping as one component of a broader workflow that links cyber incidents to MITRE ATT&CK techniques, CIS Controls and quantifiable metrics through the Cyber Catalog.

---

22 Rafiey, P. & Namadchian, A. (2025). Mapping Vulnerability Description to MITRE ATT&CK Framework by LLM. Advances in Artificial Intelligence and Machine Learning. 2025;5(3):243.
23 Rahman, F., Halim, S. M., Singhal, A., & Khan, L. (2024). ALERT: A Framework for Efficient Extraction of Attack Techniques from Cyber Threat Intelligence Reports Using Active Learning. In DBSEC 2024: Data and Applications Security and Privacy, 203–220.
24 Xu, M., Wang, H., Liu, J., Lin, Y., Xu, C., Liu, Y., Lim, H. W., & Dong, J. S. (2024). IntelEX: A LLM-driven Attack-level Threat Intelligence Extraction Framework. [Preprint]. arXiv. https://arxiv.org/html/2412.10872v1.
25 MITRE Center for Threat-Informed Defense. (2023). Threat Report ATT&CK Mapper (TRAM). https://ctid.mitre.org/projects/threat-report-attck-mapper-tram/.

The proposed approach also differs from purely generative LLM-based methods. Rather than relying on an LLM at inference time to generate or justify ATT&CK mappings, this study fine-tunes a sentence-transformer model to perform reproducible semantic matching between incident descriptions and technique descriptions. This provides a lightweight and repeatable mapping component that can be integrated into operational workflows. Nevertheless, direct comparison with recent prompt-based, retrieval-augmented and LLM-assisted ATT&CK mapping approaches remains an important direction for future work.

2.5 Contrastive Learning and Loss Functions

Contrastive learning frameworks, which learn representations by contrasting positive pairs against negative examples, have shown promise for semantic similarity tasks. MultipleNegativesRankingLoss (MNRL) is a widely adopted loss function that leverages in-batch negatives to improve training efficiency[26]. However, MNRL can introduce false negatives when the same target appears multiple times in a batch, a common scenario in cyber security datasets where multiple incidents may map to the same technique.

Works have explored modifications to contrastive learning frameworks to address these challenges, including hard negative mining and duplicate-aware training procedures[27,28]. These approaches are particularly relevant for domains characterised by one-to-many or many-to-many relationships between input and output entities.

2.6 Security Control Frameworks

Various security control frameworks have been developed to provide organisations with structured guidance for implementing cyber security measures. The CIS Security Controls represent a prioritised subset of security best practices that have been validated through widespread industry adoption[29]. Mapping threat intelligence to security controls enables organisations to translate abstract threat information into concrete defensive actions.

However, limited research has explored automated approaches for linking threat intelligence, adversary techniques, and security controls into a unified framework. Our

26 Piperno, R., Bacco, L., Dell'Orletta, F., Merone, M., & Pecchia, L. (2025). Cross-lingual distillation for domain knowledge transfer with sentence transformers. Knowledge-Based Systems, 311. https://doi.org/10.1016/j.knosys.2025.113079.
27 Cui, W., Bai, L., Yang, X., & Liang, J. (2023). A new contrastive learning framework for reducing the effect of hard negatives. Knowledge-Based Systems, 260. https://doi.org/10.1016/j.knosys.2022.110121.
28 Robinson, J., Chuang, C.-Y., Sra, S., & Jegelka, S. (2021). Contrastive learning with hard negative samples. Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).
29 Centre for Internet Security. (2021). CIS Critical Security Controls, https://www.cisecurity.org/controls/v8, Accessed 2 June 2026.

work addresses this gap by providing an end-to-end pipeline from incident description to quantifiable metrics used to monitor control implementation.

3. Methodology

3.1 Overview

Our methodology comprises five stages: (1) creating the Cyber Catalog, a comprehensive knowledge base integrating security controls, adversary techniques, and metrics; (2) training data preparation through synthetic augmentation and quality assessment; (3) model selection and architecture configuration; (4) fine-tuning with a modified loss function to address data characteristics; and (5) comprehensive evaluation using multiple metrics and baseline comparisons.

3.2 Cyber Catalog Development

3.2.1 Motivation and Design Principles

A fundamental challenge in operational cyber security is the disconnection between threat intelligence, defensive controls, and measurable outcomes. Security teams typically work with three distinct knowledge domains: (1) threat intelligence describing adversary behaviours (e.g., MITRE ATT&CK), (2) control frameworks prescribing defensive measures (e.g., CIS Controls), and (3) metrics for assessing control effectiveness. These domains are rarely integrated, forcing practitioners to manually bridge gaps between threat observations and defensive actions.

To address this fragmentation, we created the Cyber Catalog, a unified knowledge base that systematically connects these three domains. The Cyber Catalog serves as a resource that enables organisations to map observed threats (ATT&CK techniques) into relevant defensive controls (CIS Controls), assess control effectiveness using standardised and quantifiable metrics, and support risk-based prioritisation of security investments.

3.2.2 Structure and Components

The Cyber Catalog comprises three interconnected layers. The foundation layer consists of the 18 CIS Critical Security Controls (v8), which represent prioritised cyber security best practices developed through community consensus. Each control addresses a specific security domain (e.g., asset management, access control, incident response) and comprises multiple safeguards that prescribe concrete implementation actions. The second layer incorporates MITRE ATT&CK techniques representing documented adversary behaviours, with each technique mapped bidirectionally to relevant CIS controls, establishing explicit relationships between threats and defensive measures. These mappings enable security teams to determine which controls mitigate specific techniques and, conversely, which techniques a given control addresses. The third layer associates each security control with quantifiable metrics that enable objective assessment

and monitoring, with all metrics adhering to SMART criteria in line with established best practices in security metrics selection.

3.2.3 Mapping Methodology

The mapping methodology for CIS Controls to MITRE ATT&CK Techniques involves aligning each Safeguard with specific MITRE ATT&CK techniques and sub-techniques to illustrate how defensive actions can mitigate or detect adversarial behaviours. This mapping is done at a granular level, focusing on individual Safeguards rather than broad Controls, and uses the MITRE ATT&CK v8.2 Enterprise framework. This version was selected because the CIS Controls v8 Master Mapping used in this study is explicitly aligned with MITRE Enterprise ATT&CK v8.2. Using the same ATT&CK version ensured consistency between the source CIS-to-ATT&CK mapping, the Cyber Catalog structure, and the subsequent incident-to-technique mapping process. The goal is to help organisations implement threat-informed defence strategies by showing which security practices counter known attack patterns, thereby supporting prioritisation, gap analysis, and operational planning in cyber security programmes. This resource was adapted from[30].

Additionally, the CIS Controls Assessment Specification (CAS) is designed to help organisations consistently measure the implementation of each CIS Safeguard. It distinguishes between verifying whether a Safeguard has been implemented and assessing how well it has been implemented, with the specification focusing on the former. The approach emphasises what to measure, not how to measure, allowing flexibility across diverse technologies. Each Safeguard is broken down into structured components: assumptions, required inputs, operations to perform on those inputs, resulting measures, and metrics derived from those measures. This structure enables adopters to uniformly assess Safeguard implementation while accommodating technological variability[31].

Whilst the CIS Controls to MITRE ATT&CK mapping is distributed in a structured CSV format that readily supports automated analysis, the CIS Controls Assessment Specification (CAS) is accessible only via the CIS website and is not available as a downloadable structured dataset. To address this, we undertook a systematic manual extraction process, aligning each CAS component with its corresponding CIS Control, thereby enabling integrated assessment and analytical consistency across the Cyber Catalog.

30 Centre for Internet Security. CIS Controls v8 Master Mapping to MITRE Enterprise ATT&CK v82. https://www.cisecurity.org/insights/white-papers/cis-controls-v8-master-mapping-to-mitre-enterprise-attck-v82, Accessed 2 June 2026.
31 Centre for Internet Security. CIS Controls Assessment Specification. https://www.cisecurity.org/controls/cis-controls-assessment-specification, Accessed 2 June 2026.

3.3 Training Dataset Preparation

The dataset construction process followed a sequential pipeline comprising five stages: (1) extraction of cyber incident records from the European Repository of Cyber Incidents (EuRepoC); (2) manual mapping of incident descriptions to MITRE ATT&CK techniques; (3) generation of synthetic paraphrases for each manually labelled incident; (4) quality assurance and filtering of the generated examples; and (5) partitioning into training, validation and test sets for model development and evaluation.

3.3.1 Base Dataset

Each incident was subjected to a manual annotation process in which the available incident description was reviewed to identify adversary behaviours corresponding to MITRE ATT&CK techniques. The mapping considered the reported attack vector, adversary action, affected asset or environment, observed outcome, and any technical indicators described in the EuRepoC record[32]. Technique assignments were made only where the incident description provided sufficient evidence to support a plausible correspondence with a specific ATT&CK technique. Incidents that lacked sufficient behavioural or technical detail for reliable mapping were excluded from the labelled dataset. This expert annotation process yielded 762 incident-technique pairs, forming the base dataset.

The manual mapping was performed by the author, who possesses seven years of experience in cyber security, with comprehensive knowledge of the MITRE ATT&CK taxonomy and substantial expertise in analysing real-world attack patterns, thereby ensuring the reliability and validity of the annotations. The initial dataset, whilst high-quality, was insufficient in scale for effectively fine-tuning. Deep learning models, particularly those based on transformer architectures, typically require thousands to tens of thousands of training examples to achieve optimal performance[33]. This limitation necessitated a data augmentation strategy.

The manually labelled dataset was treated as the ground-truth seed dataset. Synthetic data generation was used only to create linguistically diverse variants of these manually labelled incidents; it was not used to create new technique labels independently of the original human annotation.

3.3.2 Synthetic Data Augmentation

Advanced Generative Pre-Trained Transformer (GPT), such as GPT-5 series, demonstrate superior performance in synthetic data generation tasks compared to earlier transformer models and alternative generative approaches like Generative Adversarial

32 European Repository of Cyber Incidents. EuRepoC database, https://eurepoc.eu/database, Accessed 2 June 2026.

33 Kaya, M., & Bilge, H. Ş. (2019). Deep metric learning: A survey. In Symmetry (Vol. 11, Issue 9). MDPI AG. https://doi.org/10.3390/sym11091066.

Networks (GANs). This improvement is attributed to enhanced model capacity, training methodologies, and contextual representation capabilities in latest transformer-based designs[34].

Therefore, to address the data scarcity challenge, we instructed a large language model (GPT-5) to generate synthetic incident descriptions. For each of the 762 original incidents, we instructed the model to generate 100 semantically similar but lexically distinct incident descriptions. This prompt-based augmentation approach was effective for creating diverse training data whilst preserving semantic content[35].

The prompt instructed the model to preserve the core attack behaviour, attack vector, adversary action and technical meaning of the original incident while varying surface-level linguistic expression. The generation process encouraged synonym replacement, sentence restructuring, variation in active and passive constructions, and phrase-level paraphrasing. Entity randomisation was also applied to reduce overfitting to specific named entities, including organisation names, countries, regions, malware names, threat actor names and temporal references. However, the model was instructed not to alter the underlying adversary behaviour or the ATT&CK technique implied by the incident.

To enhance the generalisability across diverse real-world scenarios, we implemented systematic entity randomisation. This process involved varying specific entities throughout the synthetic incident descriptions, including geographic references such as countries and regions, organisation names, malware and threat actor designations, temporal references encompassing dates and time periods. This randomisation ensured that the model would not overfit to specific named entities present in the original dataset. Moreover, we enforced strict constraints to ensure semantic preservation of critical incident characteristics[34]. Specifically, we maintained consistency in attack vectors and techniques employed. The augmentation process generated 76,200 synthetic incident descriptions (762 incidents times 100 augmentations), yielding a more diverse training corpus.

3.3.3 Quality Assurance and Filtering

The augmented dataset underwent quality assurance procedures: Cyber security experts manually reviewed the augmented dataset to verify accuracy and relevance. This review process identified and removed 1,214 duplicate or low-quality entries, resulting in 74,986 high-quality incident-technique pairs. Furthermore, we employed BERTScore, a reference-based evaluation metric that leverages contextual embeddings from BERT, to

34 Sufi, F. (2024). Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. Information (Switzerland), 15(2). https://doi.org/10.3390/info15020099.
35 Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., & Joty, S. (2024). Data Augmentation using Large Language Models: Data Perspectives, Learning Paradigms and Challenges. [Preprint]. https://doi.org/10.48550/arXiv.2403.02990.

assess semantic similarity between original and augmented incidents. BERTScore is superior to n-gram-based metrics for evaluating semantic similarity because it captures meaning rather than merely surface-level lexical overlap[36].

BERTScore computes precision, recall, and F1 scores by comparing token-level embeddings between reference and candidate sentences using cosine similarity. For our dataset, we established a threshold of 0.75 for the F1 score, below which augmented examples were excluded. The lowest scores observed in our filtered dataset are:

- Precision: 0.807 (indicating high relevance of generated content).
- Recall: 0.756 (indicating substantial capture of original content).
- F1: 0.781 (indicating strong overall semantic alignment).

These metrics demonstrate that our augmented dataset maintains high semantic fidelity to the original incidents whilst introducing beneficial lexical diversity.

3.3.4 Hard Negative Mining

The key challenge in our training data is the one-to-many relationship between techniques and incidents where multiple different incidents may map to the same ATT&CK technique. When using contrastive loss functions such as MNRL, this characteristic can introduce false negatives if multiple positive pairs sharing the same technique appear in the same training batch[37]. To address this and improve model discrimination, we implemented a hard negative mining strategy. Hard negatives are examples that are semantically similar to the positive example but represent incorrect mappings. Including such examples forces the model to learn finer-grained distinctions[38].

The hard negative mining procedure followed a structured similarity-based selection process. First, each ATT&CK technique description was normalised by lowercasing, removing punctuation, removing stop words, and reducing lexical variation through stemming. Second, pairwise similarity was computed between each technique and every other distinct technique using two complementary measures: stemmed word-level Jaccard similarity to capture lexical overlap, and character 4-gram overlap to capture morphological and phrase-level similarity. These two scores were combined into a single technique-similarity score. Third, for each positive technique t, the most similar but distinct technique $u \neq t$ was selected as the hard negative. The selected hard negative was then associated with each training instance containing t as the positive technique.

GPT-5 was used to support the organisation and generation of this workflow; however, the hard-negative selection criteria were based on the explicit normalisation and

36 Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1904.09675.
37 Cui, W., Bai, L., Yang, X., & Liang, J. (2023). A new contrastive learning framework for reducing the effect of hard negatives. Knowledge-Based Systems, 260. https://doi.org/10.1016/j.knosys.2022.110121.
38 Robinson, J., Chuang, C.-Y., Sra, S., & Jegelka, S. (2021). Contrastive learning with hard negative samples. Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).

similarity steps described above. This ensured that hard negatives represented semantically close but incorrect ATT&CK techniques, requiring the model to learn fine-grained distinctions between related adversary behaviours.

Therefore, this instructional approach ensures that GPT-5 generates training examples that compel the model to learn fine-grained discriminations between genuinely similar but distinct techniques, rather than relying on coarse-grained pattern recognition that might conflate related but separate attack methodologies.

The final training dataset contains 74,986 incident-technique pairs: 59,987 training set (80%), 7,499 validation set (10%), and 7,499 test set (10%). The test set was held out entirely during training and validation, providing an unbiased estimate of model performance on unseen data.

3.4 Model Architecture

3.4.1 Base Model Selection

We selected all-mpnet-base-v2 as our base model for fine-tuning. This model is a sentence transformer based on the MPNet (Masked and Permuted Pre-training for Language Understanding) architecture, which combines the advantages of BERT and XLNet whilst mitigating their respective limitations[39].The all-mpnet-base-v2 model generates 768-dimensional fixed-length sentence embeddings that capture semantic meaning and contextual information. Previous comparative studies have demonstrated that all-mpnet-base-v2 achieves statistically superior performance compared to other popular open-source sentence transformers across diverse semantic similarity benchmarks[39].

3.4.2 Framework Selection

We implemented the pipeline using PyTorch, an open-source deep learning framework renowned for flexibility, modularity, and dedicated support for research applications. PyTorch integrates seamlessly with the Sentence-Transformers library, which provides high-level APIs for training and evaluating sentence transformers[40]. Comparative studies have demonstrated that PyTorch achieves superior performance to TensorFlow in terms of training accuracy, precision, recall, and loss convergence for similar tasks. Furthermore, the sentence-transformers library is natively built on PyTorch, making it the natural choice for our implementation[41].

---

39 Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 33, 16145–16155.
40 Hussain, S. A., Kohli, R., Zahoor, S., & Sofi, S. A. (2025). Transforming the GUI Landscape: Harnessing the Power of MPNet base v2 Sentence Transformers. Procedia Computer Science, 259, 1809–1816. https://doi.org/10.1016/j.procs.2025.04.136.
41 Stanescu, L., & Dinu, G. (2023). TensorFlow vs. PyTorch in Classifying Medical Images - preliminary results. 2023 27th International Conference on System Theory, Control and Computing, ICSTCC 2023 - Proceedings, 448–453. https://doi.org/10.1109/ICSTCC59206.2023.10308472.

3.5 Loss Function

3.5.1 Multiple Negatives Ranking Loss

For semantic similarity tasks, the loss function must encourage the model to produce embeddings where similar sentences are close together in the embedding space whilst dissimilar sentences are far apart. MultipleNegativesRankingLoss (MNRL) is a widely adopted loss function for this purpose[42,43].

MNRL works with mini-batches that consist of anchor-positive pairs. For each anchor paired with its corresponding positive example, the approach considers every remaining sentence in the batch as a negative example. The objective function minimises the negative log-likelihood of similarity scores normalised through softmax, which effectively trains the model to assign a higher ranking to the correct positive example compared to all negative examples within the batch[43].

3.5.2 Duplicate-Aware Modification

A critical limitation of MNRL in our context is the one-to-many relationship in our training data: the same technique may appear as the positive target for multiple different incidents. If two pairs sharing the same technique appear in the same mini-batch, MNRL will incorrectly treat one as a negative for the other, introducing false negative signals that can degrade model performance. To address this challenge, we implemented a duplicate-aware modification using the NoDuplicatesDataLoader. This component ensures that no sentence string appears more than once within any given training batch. By eliminating within-batch duplicates of positive examples, we prevent the false negative problem whilst maintaining the efficiency benefits of in-batch negative sampling[44].

The duplicate-aware approach, combined with hard negative mining, provides a robust training signal that encourages fine-grained discrimination whilst avoiding penalisation of legitimate positive examples.

3.6 Training Configuration

3.6.1 Hyperparameters

We employed standard hyperparameter values that have been validated across various sentence transformer fine-tuning tasks45, including a learning rate of 2×10^{-5},

42 Baek, S., Park, C. Y., & Jung, W. (2025). Automated safety risk management guidance enhanced by retrieval-augmented large language model. Automation in Construction, 176. https://doi.org/10.1016/j.autcon.2025.106255.

43 Piperno, R., Bacco, L., Dell'Orletta, F., Merone, M., & Pecchia, L. (2025). Cross-lingual distillation for domain knowledge transfer with sentence transformers. Knowledge-Based Systems, 311. https://doi.org/10.1016/j.knosys.2025.113079.

44 Martellozzo, S. (2024). Integrating Semantic and Keyword Search: A Transformer-Based Approach for Content Discovery. [Master's thesis, Politecnico di Milano].

45 Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 33, 16145–16155.

a batch size of 16, and a total of 10 epochs. The training process incorporated a warmup period equivalent to 10% of the total training steps, with evaluation performed consistently after each epoch. This learning rate was selected because it provides stable optimisation without the risk of overshooting, while the batch size of 16 was chosen to represent a balance between ensuring gradient stability and maintaining computational efficiency within our hardware constraints.

3.6.2 Computational Environment

Training was conducted on a workstation running Ubuntu 24.04 LTS, with Python version 3.12.3, Sentence-Transformers library version 5.1.2, and PyTorch version 2.9.0. The hardware used was an NVIDIA Quadro RTX 5000 GPU, which features 3,072 CUDA cores, 16GB of GDDR6 memory, and is compatible with CUDA version 12.1. The entire training process, spanning 10 epochs, took approximately 46,609 seconds to complete.

3.7 Evaluation Methodology

3.7.1 Reference Similarity Scores

Each evaluation instance consisted of an incident description, a candidate MITRE ATT&CK technique description, and a reference similarity score. The reference score represented the expected semantic correspondence between the incident and the candidate technique on a normalised [0,1] scale. The manually annotated incident-technique pairs formed the seed reference set, with the assigned ATT&CK technique treated as the relevant positive mapping. Synthetic paraphrases inherited the ATT&CK technique label and reference score from their corresponding manually annotated source incident, because synthetic generation was used only to introduce linguistic variation and not to create new ground-truth labels.

For contrastive learning and evaluation, hard-negative technique candidates were associated with lower reference similarity values to reflect that they were semantically related but incorrect mappings. The final dataset therefore contained the positive ATT&CK technique mapping, an associated reference confidence/similarity score, and a hard-negative technique with its corresponding similarity value. These values were used as the ground-truth reference scores for computing Spearman correlation, Pearson correlation, MAE and MSE.

3.7.2 Evaluation Metrics

To assess model performance, we employed both correlation-based and error-based metrics. The metrics were chosen to reflect the dual nature of the task. First, the model must rank candidate ATT&CK techniques according to their relevance to a given incident description; Spearman correlation was used to assess ordering consistency. Second, the predicted similarity scores should remain numerically aligned with the ground-truth scores; therefore, Pearson correlation was included as a complementary measure. Finally, Mean Absolute Error (MAE) and Mean Squared Error (MSE) were used

to quantify prediction error directly, with MAE indicating the average magnitude of error and MSE penalising larger deviations that could be operationally significant in cyber risk prioritisation.

Semantic textual similarity tasks are evaluated using correlation coefficients that measure the alignment between predicted and ground-truth similarity scores[46]. Spearman Correlation (ρ) measures monotonic relationship strength, focusing on ranking accuracy, and is the preferred metric for semantic similarity tasks as it captures the model's ability to correctly order pairs by similarity. Pearson Correlation (r) complements this by measuring the linear relationship strength between predicted and true scores. Following Schober & Schwarte (2018), correlation values are interpreted as very strong ($|\rho| > 0.7$), moderate ($0.5 < |\rho| \leq 0.7$), fair ($0.3 < |\rho| \leq 0.5$), or poor ($|\rho| \leq 0.3$).

In addition to correlation metrics, we employed error-based metrics to assess prediction accuracy directly[47]. MAE captures the average of absolute differences between predicted and true similarity scores, whilst MSE computes the average of squared differences, penalising larger errors more heavily than MAE and thereby providing a stricter measure of prediction consistency.

3.7.3 Baseline Models

We compared our fine-tuned model against three top general-purpose sentence transformer models: all-mpnet-base-v2, known for excellent embedding quality; all-distilroberta-v1, a distilled variant of RoBERTa optimised for efficiency; and all-MiniLM-L12-v2, a compact 12-layer model optimised for speed.

The same held-out test set was used to evaluate our fine-tuned model against these models, using identical evaluation procedures to ensure fair comparison.

4. Results

4.1 Training Performance

Training progressed smoothly over 10 epochs, with consistent improvement in validation set performance. The duplicate-aware loss function successfully mitigated the false negative problem, as evidenced by stable convergence without the oscillations that can occur when the model receives contradictory training signals.

Our fine-tuned model achieved strong correlation metrics on the test set. Both metrics reveal a strong alignment between the predicted and ground-truth similarity scores. A Spearman correlation of 0.789 surpasses the 0.7 threshold for a very strong

46 Rovetta, A. (2020). Raiders of the Lost Correlation: A Guide on Using Pearson and Spearman Coefficients to Detect Hidden Correlations in Medical Sciences. Cureus. https://doi.org/10.7759/cureus.11794.
47 Galli, C., Donos, N., & Calciolari, E. (2024). Performance of 4 Pre-Trained Sentence Transformer Models in the Semantic Query of a Systematic Review Dataset on Peri-Implantitis. Information (Switzerland), 15(2). https://doi.org/10.3390/info15020068.

correlation, indicating that the model consistently and reliably ranks incident-technique pairs by their similarity[48]. Furthermore, the higher Pearson correlation of 0.876 suggests that this relationship is not only monotonic but also closely approximates a linear trend.

4.2 Comparative Analysis

Table 1 presents the Spearman correlation performance of our fine-tuned model compared to the top baseline models, along with the delta spearman (Δρ) representing the improvement over each model.

Table 1. Spearman Correlation Comparison.

Model	**Spearman (ρ)**	**Δρ (Improvement)**
Our Fine-Tuned Model	0.7894	—
all-mpnet-base-v2	0.5852	+0.2042
all-distilroberta-v1	0.5776	+0.2118
all-MiniLM-L12-v2	0.5585	+0.2309

Our fine-tuned model outperforms all baseline models, with improvements ranging from 0.20 to 0.23 in Spearman correlation. These represent relative improvements of approximately 37% over the baseline models. Notably, even the base model (all-mpnet-base-v2) before fine-tuning achieves only moderate correlation (ρ = 0.585), demonstrating that general-purpose transformers, whilst capable, lack the domain-specific knowledge necessary for accurate mapping.

Table 2 presents the Error metrics for our fine-tuned model and all baseline models.

Table 2. Mean Absolute and Squared Error Comparison.

Model	**MAE**	**MSE**
Our Fine-Tuned Model	0.1352	0.0272
all-mpnet-base-v2	0.5281	0.2859
all-distilroberta-v1	0.4806	0.2417
all-MiniLM-L12-v2	0.3033	0.0941

Our model exhibits significantly lower error rates compared to all baseline models (see Fig. 1). The MAE of 0.1352 indicates on average, our model's predictions deviate from ground truth by only 0.1352 units on the similarity scale. This represents approximately 67% reduction in MAE compared to the baseline models.

The MSE results tell a similar story: our model achieves an MSE of 0.027, compared to 0.0941–0.2859 for baseline models. The lower MSE indicates that our model

48 Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. Anesthesia and Analgesia, 126(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864.

not only has lower average error but also produces fewer large errors, an important characteristic for practical deployment where outlier predictions can have significant consequences.



Fig. 1. MAE and MSE Comparison.

From a practical perspective, these improvements are important because the model is intended to support analyst-facing incident triage rather than operate as a purely abstract similarity benchmark. Higher rank correlation means that relevant ATT&CK techniques are more likely to appear near the top of the candidate list, reducing the effort required by analysts to identify plausible mappings. Similarly, lower MAE and MSE indicate that the model produces fewer large deviations from the reference similarity scores, which reduces the likelihood that unrelated or weakly related techniques are prioritised during downstream control selection and risk assessment.

4.3 Error Distribution Analysis

To gain deeper insight into model behaviour, we examined the distribution of absolute errors across the test set. Fig. 2 illustrates the error distributions for all models.

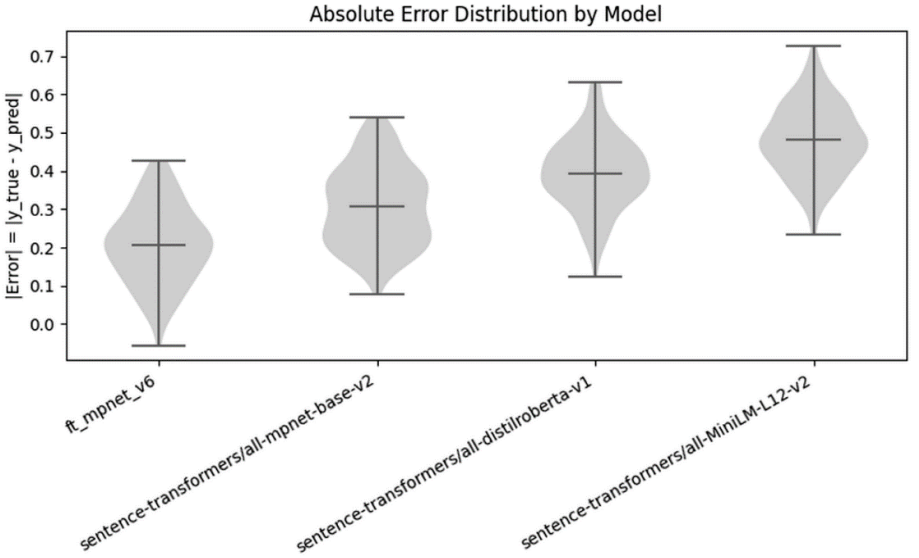


Fig. 2. Violin plot of absolute error distribution.

The key observation from the analysis of model performance distributions is that our model exhibits the lowest median error, as indicated by the thick horizontal bar within the violin plot, thereby confirming superior average-case performance relative to baseline approaches.

Moreover, the distribution for our model is notably narrow and tightly concentrated near zero, indicating consistent accuracy across the entire test set rather than performance that varies between different instances. Importantly, our model exhibits minimal probability density in high-error regions, whereas baseline models demonstrate non-negligible probability mass at large error values, suggesting that these baselines are susceptible to occasional severe mispredictions. Collectively, this narrow error distribution demonstrates that our model is not only more accurate on average but also exhibits greater stability and predictability in its behaviour, a critical characteristic for deployment in operational cyber security contexts where reliability is paramount.

5. Discussion

5.1 Interpretation of Results

The experimental results demonstrate that domain-specific fine-tuning yields substantial improvements over general-purpose sentence transformers for the task of mapping cyber incidents to MITRE ATT&CK techniques. The Spearman correlation improvement of approximately 0.2 over baseline models represents a meaningful practical enhancement: the difference between a model that provides useful similarity rankings ($\rho \approx 0.59$) and one that provides reliable, actionable rankings ($\rho \approx 0.79$).

Furthermore, the error analysis provides additional insight into model behaviour. The three-fold reduction in MAE compared to top baseline models indicates that our fine-tuned model's predictions are closer to ground truth. More importantly, the narrow error distribution demonstrates consistency, a critical property for building trust in automated systems that support decision-making.

5.2 Impact of Training Data Quality

The observed performance gains can be attributed to several methodological choices. First, LLM-based data augmentation expanded the manually labelled dataset from 762 incident-technique pairs to 74,986 pairs while preserving semantic quality. BERTScore ensured that only high-quality examples are retained, preventing the model from learning spurious patterns from poorly generated data. Moreover, the inclusion of hard negatives, semantically similar but incorrect techniques, forced the model to learn fine-grained distinctions rather than merely broad categorical differences. The modification to address false negatives arising from one-to-many mappings prevented the model from receiving contradictory training signals. This contributed to more stable and effective learning.

5.3 Practical Applications

The fine-tuned model, combined with the Cyber Catalog, enables practical applications as outlined below.

Security analysts can use the integrated framework to automatically map incidents to relevant ATT&CK techniques and subsequently to applicable CIS controls. This end-to-end automation dramatically reduces the time required for incident triage and response planning. The fine-tuned model identifies relevant ATT&CK techniques from incident descriptions, enabling rapid classification even during high-volume incident scenarios. This acceleration in speed is advantageous in organisations facing resource limitations and having finite capacity.

By mapping incidents to techniques and subsequently to CIS controls via the Cyber Catalog, the framework enables organisations to identify gaps in their security posture. If incidents consistently map to techniques for which relevant controls are not implemented, this signals a prioritisation opportunity for investments.

Moreover, the Cyber Catalog's integration of metrics enables objective, quantifiable assessment of control effectiveness. Rather than relying on subjective judgement or compliance checklists, organisations can track metrics over time to determine whether implemented controls are reducing relevant incidents and achieving desired security outcomes. This evidence-based approach supports resource allocation decisions and demonstrates security programme value to stakeholders.

The automated mapping enables data-driven risk analysis. By aggregating incident patterns over time and correlating them with control implementation status and metrics, organisations can identify which techniques pose the highest threat to their environment, which controls provide the highest risk reduction, and where to allocate resources for maximum impact. The structured mappings in the Cyber Catalog facilitate compliance reporting by connecting implemented controls to potential threats and measurable outcomes.

Furthermore, the Cyber Catalog serves as an educational resource for less experienced security practitioners, helping them understand relationships between threats, controls, and metrics. This supports organisations with limited expertise.

5.4 Addressing the Research Questions

Sentence transformer models, when fine-tuned on domain-specific data, can facilitate accurate automated correspondence between cyber incidents and security controls via MITRE ATT&CK techniques. With respect to RQ1, key challenges include: (1) obtaining sufficient high-quality training data, which we addressed through synthetic data augmentation; (2) handling one-to-many mappings, which we addressed through duplicate-aware training; and (3) achieving fine-grained discrimination between similar techniques, which we addressed through hard negative mining. Objective metrics based on model prediction accuracy (MAE & MSE) and rank correlation (Spearman ρ) provide quantifiable assessments of model performance.

Regarding RQ2, general-purpose transformer models achieve moderate performance ($\rho \approx 0.59$) on this task, indicating that they capture semantic patterns relevant to incident-technique mapping. However, domain-specific fine-tuning (our fine-tuned model) yields substantial improvements ($\Delta\rho \approx 0.21$), elevating performance to the very strong correlation regime.

To address RQ3, the automated mappings enable decision-support capabilities, such as prioritisation of control implementation based on observed incident patterns, objective assessment of control effectiveness through metrics, and identification of security posture gaps.

5.5 Limitations

The cyber security landscape evolves rapidly, with new attack techniques emerging regularly. Our model was trained on historical incidents and may require periodic retraining to maintain accuracy as the threat landscape changes. Moreover, the model was trained exclusively on English incident descriptions. We did not consider incidents described in other languages; this would require multilingual models or translation pipelines.

Although the dataset was divided into separate training, validation and test sets, synthetic paraphrase generation can introduce a risk that closely related examples may influence performance estimates if augmented samples derived from the same source incident are not grouped consistently. Future work will therefore extend the evaluation using grouped splitting and grouped cross-validation, where each original EuRepoC incident and all synthetic paraphrases derived from it are kept within the same split or fold.The statistical analysis of model performance is also limited. The current study reports correlation-based and error-based metrics on the held-out test set, which allows comparison between the fine-tuned model and baseline sentence-transformer models. However, inferential statistical testing, confidence intervals or bootstrap-based significance analysis were not conducted. Future work will include statistical significance testing and confidence interval estimation to assess the robustness of the observed performance differences across repeated splits or grouped cross-validation folds.

Further validation is required under operational conditions and on independent real-world incident datasets. Although the current evaluation demonstrates the potential of the proposed framework, future work will assess its generalisability using external incident corpora and analyst-facing workflows containing naturally occurring, incomplete and noisy incident descriptions.The successful integration of these models into operational workflows necessitates the establishment of robust governance protocols. Specifically, organisations should implement confidence thresholds below which automated mappings are automatically flagged for manual adjudication.

The ATT&CK annotation process also has limitations. Although technique labels were assigned through expert review, mapping public incident descriptions to ATT&CK techniques can involve subjectivity, especially when reports provide limited behavioural

or technical detail. Future work will strengthen label reliability through multi-annotator validation, adjudication of disagreements, and inter-annotator agreement analysis.

A further limitation is that the current study did not include an ablation analysis of the individual methodological components. Consequently, while the final model demonstrates improved performance over the selected baselines, the separate contribution of synthetic data augmentation was not quantified independently. Future work will conduct ablation experiments comparing models trained with and without each component to determine their individual effect on ranking accuracy, prediction error, and robustness.

6. CONCLUSION

This research has presented a comprehensive framework for automating the mapping of cyber incidents to MITRE ATT&CK techniques and subsequently to actionable security controls using a fine-tuned sentence transformer model. A central contribution of this work is the Cyber Catalog, a novel knowledge base that systematically integrates CIS Critical Security Controls, MITRE ATT&CK techniques, and quantifiable metrics. This integrated resource addresses a fundamental gap in operational cyber security by enabling organisations to translate threat intelligence into defensive actions and measurable outcomes. Through rigorous methodology encompassing data augmentation with quality controls, duplicate-aware contrastive learning, and hard negative mining, we achieved substantial performance improvements over top baseline models. Our fine-tuned model attained a Spearman correlation of 0.789 and Pearson correlation of 0.876, representing improvements of approximately 0.21 over baseline models.

Furthermore, our model exhibited significantly lower prediction errors (MAE: 0.135 & MSE: 0.027), these metrics demonstrated three-fold reductions in MAE and MSE compared to the top baseline models, with a narrow error distribution confirming consistent, reliable predictions across diverse incident types. The framework addresses critical operational challenges in cyber risk management: accelerating incident triage, enabling security control gap identification, facilitating objective control effectiveness assessment through quantifiable metrics, and supporting data-driven risk management. By bridging the gap between unstructured threat intelligence and structured security controls with objective metrics, this work provides practical value to organisations, particularly resource-constrained SMEs, seeking to strengthen their cyber security posture through evidence-based decision-making.

Future work will integrate the fine-tuned model into an open-source host intrusion detection system through a specialised plugin that uses the Cyber Catalog for near-real-time analysis and response. Another direction is to expand the Cyber Catalog by incorporating mappings to NIST Special Publication 800-53 controls. This would broaden the framework's applicability and support organisations that need to align their cyber security posture with diverse regulatory and compliance requirements.

By reducing the time required for incident analysis, the proposed framework can help narrow the capability gap between resource-rich and resource-constrained organisations. This is particularly relevant for SMEs, which often lack the cyber security resources available to larger enterprises. Overall, the study supports a shift from reactive, manual incident analysis toward more proactive and evidence-based cyber risk management.

DATA AVAILABILITY

The datasets used in this study are publicly available and can be downloaded from the following link: https://zenodo.org/records/17670439.

BIBLIOGRAPHY

1. Baek, S., Park, C. Y., & Jung, W. (2025). Automated safety risk management guidance enhanced by retrieval-augmented large language model. *Automation in Construction*, *176*. https://doi.org/10.1016/j.autcon.2025.106255.
2. Bayer, M., Frey, T., & Reuter, C. (2023). Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. *Computers and Security*, *134*. https://doi.org/10.1016/j.cose.2023.103430.
3. Centre for Internet Security. (2021). CIS Critical Security Controls Version 8. Retrieved 24 November 2025, from https://www.cisecurity.org/controls/v8.
4. Centre for Internet Security. (2025). CIS Controls Assessment Specification. Retrieved 24 November 2025, from https://www.cisecurity.org/controls/cis-controls-assessment-specification.
5. Centre for Internet Security. (2024). CIS Controls v8 Master Mapping to MITRE Enterprise ATT&CK v82. Retrieved 24 November 2025, from https://www.cisecurity.org/insights/white-papers/cis-controls-v8-master-mapping-to-mitre-enterprise-attck-v82.
6. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 1–14.
7. Cui, W., Bai, L., Yang, X., & Liang, J. (2023). A new contrastive learning framework for reducing the effect of hard negatives. *Knowledge-Based Systems*, *260*. https://doi.org/10.1016/j.knosys.2022.110121.
8. Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., & Joty, S. (2024). Data Augmentation using Large Language Models: Data Perspectives, Learning Paradigms and Challenges. [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2403.02990.
9. European Repository of Cyber Incidents. (2024, May 2). *EuRepoC database*. https://eurepoc.eu/database/.

10. Galli, C., Donos, N., & Calciolari, E. (2024). Performance of 4 Pre-Trained Sentence Transformer Models in the Semantic Query of a Systematic Review Dataset on Peri-Implantitis. *Information (Switzerland)*, *15*(2). https://doi.org/10.3390/info15020068.
11. Georgescu, T. M., Iancu, B., & Zurini, M. (2019). Named-entity-recognition-based automated system for diagnosing cybersecurity situations in IoT networks. *Sensors (Switzerland)*, *19*(15). https://doi.org/10.3390/s19153380.
12. Hubbard, D. W., & Seiersen, R. (2023). *How to measure anything in cybersecurity risk* (2nd ed.). Wiley.
13. Hussain, S. A., Kohli, R., Zahoor, S., & Sofi, S. A. (2025). Transforming the GUI Landscape: Harnessing the Power of MPNet base v2 Sentence Transformers. *Procedia Computer Science*, *259*, 1809–1816. https://doi.org/10.1016/j.procs.2025.04.136.
14. Kaya, M., & Bilge, H. Ş. (2019). Deep metric learning: A survey. In *Symmetry* (Vol. 11, Issue 9). MDPI AG. https://doi.org/10.3390/sym11091066.
15. Martellozzo, S. (2024). Integrating Semantic and Keyword Search: A Transformer-Based Approach for Content Discovery. [Master's thesis, Politecnico di Milano].
16. MITRE. (2023). MITRE ATT&CK. Retrieved 24 November 2025, from https://attack.mitre.org/.
17. MITRE Center for Threat-Informed Defense. (2023). Threat Report ATT&CK Mapper (TRAM). Retrieved 24 November 2025, from https://ctid.mitre.org/projects/threat-report-attck-mapper-tram/.
18. Mohasseb, A., Aziz, B., Jung, J., & Lee, J. (2019). Predicting Cyber Security Incidents Using Machine Learning Algorithms: A Case Study of Korean SMEs. *In Proceedings of the 5th International Conference on Information Systems Security and Privacy*, 230-237.
19. Pavlyshenko, B., & Stasiuk, M. (2025). Semantic similarity analysis using transformer-based sentence embeddings. *Electronics and Information Technologies*, *30*. https://doi.org/10.30970/eli.30.4.
20. Piperno, R., Bacco, L., Dell'Orletta, F., Merone, M., & Pecchia, L. (2025). Cross-lingual distillation for domain knowledge transfer with sentence transformers. *Knowledge-Based Systems*, *311*. https://doi.org/10.1016/j.knosys.2025.113079.
21. Rafiey, P. & Namadchian, A. (2025). Mapping Vulnerability Description to MITRE ATT&CK Framework by LLM. *Advances in Artificial Intelligence and Machine Learning*. 2025;5(3):243.
22. Rahman, F., Halim, S. M., Singhal, A., & Khan, L. (2024). ALERT: A Framework for Efficient Extraction of Attack Techniques from Cyber Threat Intelligence Reports Using Active Learning. *In DBSEC 2024: Data and Applications Security and Privacy*, 203–220.
23. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing, 3982–3992. https://doi.org/10.18653/v1/D19-1410.

24. Robinson, J., Chuang, C.-Y., Sra, S., & Jegelka, S. (2021). Contrastive learning with hard negative samples. *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.
25. Rovetta, A. (2020). Raiders of the Lost Correlation: A Guide on Using Pearson and Spearman Coefficients to Detect Hidden Correlations in Medical Sciences. *Cureus*. https://doi.org/10.7759/cureus.11794.
26. Savola, R. M. (2013). Quality of security metrics and measurements. *Computers and Security*, *37*, 78–90. https://doi.org/10.1016/j.cose.2013.05.002.
27. Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864.
28. Sherif, E. (2025). Task-specific dataset for Operationalising Cyber Risk Management Using AI: Connecting Cyber Incidents to MITRE ATT&CK Techniques, Security Controls, and Metrics. In Operationalising Cyber Risk Management Using AI: Connecting Cyber Incidents to MITRE ATT&CK Techniques, Security Controls, and Metrics [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17670439.
29. Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems, 33*, 16145–16155. https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.
30. Stanescu, L., & Dinu, G. (2023). TensorFlow vs. PyTorch in Classifying Medical Images - preliminary results. *2023 27th International Conference on System Theory, Control and Computing, ICSTCC 2023 - Proceedings*, 448–453. https://doi.org/10.1109/ICSTCC59206.2023.10308472.
31. Stankevičius, L., & Lukoševičius, M. (2024). Extracting Sentence Embeddings from Pretrained Transformer Models. *Applied Sciences (Switzerland)*, *14*(19). https://doi.org/10.3390/app14198887.
32. Sufi, F. (2024). Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. Information (Switzerland), 15(2). https://doi.org/10.3390/info15020099.
33. Xu, M., Wang, H., Liu, J., Lin, Y., Xu, C., Liu, Y., Lim, H. W., & Dong, J. S. (2024). IntelEX: A LLM-driven Attack-level Threat Intelligence Extraction Framework. [Preprint]. arXiv. https://arxiv.org/html/2412.10872v1.
34. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & A.rtzi, Y. (2020). BERTScore: Evaluating text generation with BERT. [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1904.09675.