Enhancing RMF and ATT&CK Mapping Accuracy through Sentence-BERT and Mitigation Parameters Integration

Hanhee Lee , Sukjoon Yoon , Yunkyung Lee , Jiwon Kang *

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1795.v1

Keywords: RMF controls; ATT&CK; sentence-BERT; semantic mapping; mitigation expansion

 doi:10.20944/preprints202602.1795.v1

Article

Enhancing RMF and ATT&CK Mapping Accuracy Through Sentence-BERT and Mitigation Parameters Integration

Hanhee Lee [1,2], Sukjoon Yoon [1,2], Yunkyung Lee [3] and Jiwon Kang [1,2,*]

[1] Department of Computer Engineering, Sejong University, Seoul 05006, Republic of Korea

[2] Defense AI & Cyber Convergence Research Institute, Sejong University, Seoul 05006, Republic of Korea

[3] Artificial Intelligence Computing Research Laboratory; Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea

* Correspondence: jwkang@sejong.ac.kr

Abstract

To minimize cybersecurity risks in weapon systems, the implementation of the Korean Risk Management Framework (K-RMF) has become essential. However, a significant 'strategic gap' exists between high-level RMF controls and technical MITRE ATT&CK techniques, rendering manual mapping labor-intensive. This study proposes an automated mitigation-driven pipeline that integrates Sentence-BERT (SBERT) with the structural defense relationships of the ATT&CK knowledge graph. To address the data coverage limitations of the CTID silver standard, we introduce Recall@restricted as a calibrated performance metric. Experimental evaluation demonstrated that the proposed ensemble framework achieves a Recall@restricted of 0.74, significantly outperforming baseline SBERT-only models. These findings suggest that deterministic mitigation relationships effectively complement semantic representations, providing a robust framework for aligning RMF controls with adversarial behaviors.

Keywords: RMF controls; ATT&CK; sentence-BERT; semantic mapping; mitigation expansion

1. Introduction

Ministry of National Defense (MND) has prioritized the implementation of the Korean Risk Management Framework (K-RMF) since April 2024. Within this framework, RMF control items exhibit a dependency across connectivity spanning vulnerabilities, software weaknesses, attack techniques, and defensive measures. The modern cybersecurity landscape has shifted toward sophisticated threats that can no longer be defended against by mere regulatory compliance. While organizational security investments and operations are still planned and audited based on control frameworks such as NIST SP 800-53 [1], threat intelligence and incident response are increasingly executed through adversarial behavior models like MITRE ATT&CK [2]. Consequently, there is a critical need for organizations to transition from a compliance-driven defense to a proactive posture that can respond rapidly to evolving adversarial tactics and techniques. In this environment, establishing an organic linkage between the static NIST SP 800-53 and the dynamic MITRE ATT&CK is essential for strengthening an organization's security posture. Without this connection, the security controls managed by Governance, Risk, and Compliance (GRC) teams fail to translate into detection and response activities for the Security Operations Center (SOC) [1-3]. Conversely, it becomes difficult to systematically explain which control weaknesses are exploited by the attack techniques observed by the SOC. The fundamental challenge in resolving this issue is "mapping"—deriving the relational links that define which NIST controls prevent, detect, or suppress specific ATT&CK techniques. However, direct mapping (Control → Technique) is difficult to automate due to a profound semantic gap [3].

 doi:10.20944/preprints202602.1795.v1

While NIST SP 800-53 Revision 5 defines "Requirements" for security and privacy controls for federal and critical infrastructure systems, MITRE ATT&CK categorizes and describes observed adversarial "Behaviors" into Tactics and Techniques [1,2]. NIST controls are composed of policy- and procedure-oriented sentences focused on "what to do," whereas ATT&CK techniques utilize action- and process-oriented descriptions of "how they do it." Even when addressing the same defensive objective, the vocabulary, sentence structure, and level of abstraction differ significantly, causing simple keyword-based matching to trigger excessive false positives. Furthermore, manual mapping by experts inevitably faces challenges related to cost, time, and subjectivity.

To bridge this gap, this study proposes an automated mapping pipeline that combines semantic embedding using Sentence-BERT (SBERT) with the reasoning capabilities of Large Language Models (LLMs) enhanced by Mitigation parameters [4]. The core innovation lies in avoiding the direct connection between controls and techniques. Instead, we utilize the "Mitigation" (Course-of-Action) objects within ATT&CK as an intermediary layer. In this structure, the Control → Mitigation link is established through semantic similarity (Probabilistic Link), while the Mitigation → Technique link leverages the deterministic "mitigates" relationship defined in the STIX knowledge graph (Deterministic) [5,14]. By restricting the scope of natural language inference—which carries inherent uncertainty—to the Control → Mitigation segment and utilizing deterministic relationships for the remainder, the overall reliability of the mapping is significantly enhanced. This study is based on the following core hypothesis:

- The semantic similarity between NIST controls and ATT&CK mitigations is clearer and more robust than the direct similarity between controls and techniques; therefore, leveraging the hard-coded relationships between mitigations and techniques will drastically improve mapping accuracy.

The remainder of this paper is organized as follows. Section 2 provides a structural overview of NIST SP 800-53 and MITRE ATT&CK, alongside an analysis of SBERT and Mitigation parameters. Section 3 details the dataset composition, preprocessing methods, and the proposed mapping pipeline architecture. Section 4 presents the experimental results of the proposed pipeline. Finally, Section 5 concludes the paper by synthesizing the findings and suggesting directions for future research.

2. Related Works

In the modern cybersecurity domain, text-based automated mapping technologies are evolving from simple keyword matching to deep-learning-based semantic similarity analysis. This section analyzes four pivotal previous studies that influenced the design of our proposed mapping pipeline.

2.1. Embedding-based Semantic Mapping Pipelines

Recent approaches have increasingly utilized BERT-family language models to perform matching between documents or sentences based on semantic similarity. Sentence-BERT (SBERT), in particular, is advantageous for large-scale comparisons as it independently embeds sentences using a Siamese architecture and then calculates cosine similarity [6-8]. However, it remains challenging to establish a definitive causal relationship—such as "Control X mitigates Technique Y" based solely on semantic similarity, and results vary significantly depending on threshold settings. While embedding models excel at proposing broad candidates, confirming final relationships requires external knowledge such as rules, graphs, or expert validation. In this study, we utilize the SBERT architecture to calculate Sim(c, m). Unlike traditional BERT models that require N x K heavy computations to compare N times controls and K times mitigations, SBERT transforms each sentence into an independent high-dimensional vector, allowing for superior computational efficiency and accuracy by calculating only the cosine similarity between vectors [6].

Wudali et al. [6] proposed the RAM framework, which maps Security Information and Event Management (SIEM) rules to ATT&CK techniques using LLM prompt chaining and context injection.

 doi:10.20944/preprints202602.1795.v1

This study achieved encouraging performance—Recall 0.75, Precision 0.52, and F1 0.62—without a separate training process. Wudali demonstrated that natural language conversion and evidence-based refinement are key to performance enhancement. However, heavy reliance on text similarity led to hallucinations and semantic confusion between similar techniques. Drawing on this multi-stage pipeline structure, our study incorporates a "Normalization–Embedding Candidate Generation–LLM Re-ranking" process tailored for the RMF domain.

Abderehman et al. [7] introduced VMTT&RP, which combines Mini LM embeddings, cosine similarity, and multi-layer perceptrons (MLP) to predict ATT&CK techniques from CVE descriptions. While it proved the possibility of automated linkage between vulnerabilities and attack techniques with a Micro-F1 of 0.68, Abderehman highlighted the difficulty in distinguishing subtle differences between similar ATT&CK techniques. Incorporating these findings, our study reflects a precision-recall balance logic by adjusting thresholds during the candidate generation phase for RMF control items.

Zhang et al. [8] proposed the VTT-LLM architecture, which injects Common Weakness Enumeration (CWE) and Common Attack Pattern Enumeration and Classification (CAPEC) knowledge through LLM fine-tuning. This approach improved tactic prediction accuracy to 85.18%, demonstrating that a chain-based knowledge injection method—linking Control → (Mitigation/Data Source) → Technique—resulted in an average performance increase of 9.24 percentage points. Although Zhang raised concerns regarding potential overlaps between evaluation and training sets, this logic provides critical insights for designing inference paths that connect RMF control items to ATT&CK techniques via mitigation strategies. Our research adopts this logical chain, utilizing RMF mitigation signals as a core element for LLM re-ranking.

Rafiey et al. [9] showed that few-shot prompting using state-of-the-art models like GPT-4o and OpenAI o1 outperformed BERT-based models, reaching a peak F1-score of 59%. Conversely, Rafiey observed that performance degrades when vulnerability descriptions are ambiguous or terminologies are inconsistent due to the limitations of keyword-based matching. To overcome these hurdles, our study integrates a closed-loop Retrieval-Augmented Generation (RAG) structure, hybrid re-ranking, and evidence span citation techniques to enhance mapping consistency. A comparative summary of existing research on automated mapping across various knowledge bases is presented in Table 1.

Table 1. Comparative analysis of automated mapping studies across various security knowledge bases.

Mapping Method	Mapping Target	Mapping Approach
MITRE CTID Mapping	RMF ↔ ATT&CK	Manual, community-based large-scale mapping
Wudali [6] NIST 800-160 v2 ↔ ATT&CK	NIST 800-160 v2 → ATT&CK	Theory-based mapping utilizing PETE analysis
Abderehman[7] RAM (Rule ATT&CK Mapper)	SIEM Rules → ATT&CK	LLM prompt chaining-based automation
CVE ↔ ATT&CK[8]	CVE Descriptions → ATT&CK Tactics	Transformer-based classification models
rcATT Tool[8]	Report Text → ATT&CK	NLP classifier-based automated extraction

2.2. NIST SP 800-53 and MITRE ATT&CK

Mapping between NIST 800-53 and ATT&CK has been pursued by various organizations and research communities. Most notably, the **Center for Threat-Informed Defense (CTID)** developed and released a mapping between NIST 800-53 controls and ATT&CK techniques based on expert consensus [3]. While these ground-truth datasets are highly practical, they possess certain limitations: (i) the cost of manual construction is substantial; (ii) the update cycle may lag behind the rapid evolution of adversarial techniques; and (iii) the mapping scope can vary depending on specific perspectives (e.g., technology-centric vs. control-centric). Academically, keyword-based or rule-

 doi:10.20944/preprints202602.1795.v1

based mapping has long been utilized. However, because control statements are primarily policy- and procedure-oriented, frequent **False Positives** occur—where the model identifies a relationship due to similar terminology despite the absence of an actual defensive correspondence. Conversely, **False Negatives** occur when the model fails to identify a relevant link because of differing terminology. This is commonly referred to in Natural Language Processing (NLP) as the "**lexical mismatch**" problem.

NIST 800-53 focuses on "**What to do.**" For instance, the **AC-2 (Account Management)** control stipulates that an organization must create, disable, and monitor accounts [1]. This resides within the domains of "Policy" and "Procedure," often without specifying the exact technical tools or the specific attacks to be thwarted. In terms of data structure, the **Control Identifier (ID)**, **Control Statement (Text)**, and **Discussion** serve as core fields. For effective NLP, extracting the semantic context from these text fields is crucial.

MITRE ATT&CK describes "**How they do it**" [2]. For example, the **Valid Accounts (T1078)** technique explains how an attacker uses compromised accounts to access a system. This resides within the domain of specific "**Actions.**" From a defensive standpoint, ATT&CK includes the concept of **Mitigations**, which refers to technical measures (e.g., **M1036 Account Use Policies**) designed to deter specific attack techniques [10].

2.3. Knowledge Graph-based Mapping and Mitigation Parameters

The MITRE ATT&CK framework, structured according to the Structured Threat Information Expression (STIX™) 2.1 standard, provides more than a simple taxonomy; it offers a relational database of adversarial behavior [5,14,19,20]. Within this schema, the relationship between a Mitigation (course-of-action) and a Technique (attack-pattern) is explicitly defined through the "mitigates" relationship type [10]. These relationships are not merely probabilistic but are "deterministic links" curated by domain experts to represent verified defensive measures against specific TTPs.

Recent research in Knowledge Graph (KG) applications for cybersecurity emphasizes that such structured relationships can significantly reduce the semantic ambiguity inherent in natural language descriptions [12, 17]. According to [17], representing security controls as nodes and their functional overlaps as edges allows for a multidimensional analysis of defense coverage. In this study, we leverage these deterministic STIX-defined links to construct a robust mapping pipeline. Unlike traditional approaches that rely solely on surface-level text similarity, which often suffers from the "polysemy problem" (where the same technical term carries different weights in policy versus implementation), our methodology separates the mapping process into two distinct layers: uncertain relationships (inferred via SBERT-based natural language similarity) and deterministic relationships (extracted from the ATT&CK/STIX graph)[16]. The integration of Mitigation parameters serves as a bridging mechanism [10]. For instance, when an RMF control (e.g., AC-2 Account Management) is evaluated, its functional objectives are decomposed into mitigation-level actions. By identifying which ATT&CK Mitigations (e.g., M1018: User Account Management) correspond to these RMF objectives, the pipeline can traverse the KG to reach the associated Techniques (e.g., T1078: Valid Accounts) with high precision. This "hybridization" of neural embeddings and graph-based logic ensures both the explainability of the mapping—by providing a clear defensive lineage—and the stability of the overall system against the noise typical of high-level policy text. Furthermore, the use of "Data Sources" and "Data Components" within the ATT&CK 10.0+ versions allows for a "Gating Technique" that validates whether a specific control actually produces the telemetry required to detect a technique. The lack of structural consistency between what a control claims to do and what a detection mechanism requires is a primary cause of false positives in automated mapping. By constraining the search space using these structural parameters, our pipeline achieves a more granular alignment than previous models that treated RMF and ATT&CK as isolated silos. This approach aligns with the "Threat-Informed Defense" philosophy, transforming static compliance checklists into dynamic, actionable intelligence.

 doi:10.20944/preprints202602.1795.v1

3. Proposed Method

Mapping Pipeline for RMF and MITRE ATT&CK via SBERT and Mitigation Parameters.

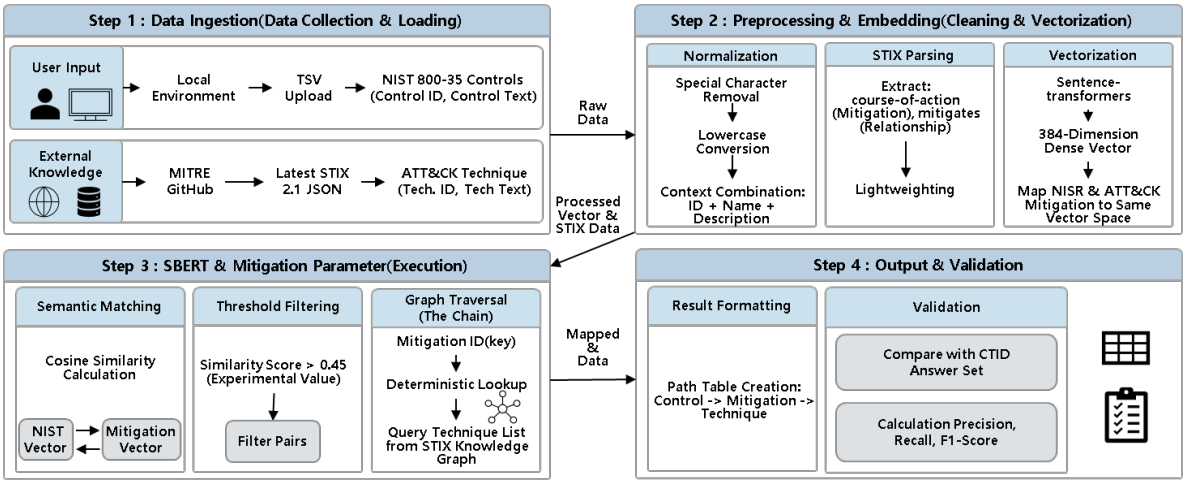

Figure 1. Architecture of the mapping pipeline for RMF and ATT&CK using SBERT and Mitigation parameters.

The RMF-to-ATT&CK mapping architecture proposed in this study consists of four primary stages: (i) Data Integration, (ii) Embedding and Candidate Generation, (iii) Mapping Decision via Mitigation Parameters, and (iv) Validation.

3.1. Problem Definition and Performance Metric Refinement

The primary objective of this study is to automatically derive the Top-K (where K=5 in this experiment) most relevant ATT&CK techniques T_k for each security control C_i. In other words, the goal is to recommend a prioritized list of attack techniques that a specific control can effectively mitigate or counteract.

To evaluate the model, we first define the ground-truth dataset. This study utilizes the control-to-technique mapping pairs provided by the Center for Threat-Informed Defense (CTID), which were established through expert consensus, as the reference set. We denote this ground-truth set as G, which is a subset of the Cartesian product of the control set C and the technique set T, containing only the pairs where a valid mapping exists. The set of results predicted by our model is denoted as G. This represents the collection of control-technique pairs that the model identifies as "relevant." A **True Positive (TP)** occurs when the model's prediction aligns with the ground truth. Mathematically, this is defined as the intersection of the predicted set and the ground-truth set : $TP = G^{\wedge} \cap G$

In essence, TP counts the number of pairs that the model correctly identified as valid mappings. The performance metrics used to evaluate this process are defined in table 2.

Table 2. Performance Metric.

Metric	Definition
Precision	Precision represents the proportion of valid linkages within the Top-K mappings proposed by the model; achieving high precision is essential to mitigating the cognitive burden and alert fatigue of security operators.
Recall	Recall represents the proportion of valid ground-truth mappings that the model successfully identifies within its Top-K candidates. This metric is critical for identifying potential security gaps, as it measures the model's ability to ensure that no essential defensive links are overlooked.
F1-score	By calculating the harmonic mean of Precision and Recall, the F1-score provides a singular, balanced indicator of a model's mapping accuracy and retrieval capability. This ensures a rigorous assessment of the model's effectiveness in aligning RMF controls with ATT&CK techniques under identical experimental conditions.

 doi:10.20944/preprints202602.1795.v1

The primary objective of this study is to automatically derive the Top-K (where K=5) most relevant ATT&CK techniques T_k for each security control C_i. To evaluate this process, we first define the standard metrics—Precision, Recall, and F1-score—as outlined in Table 3. However, a critical challenge arises: the CTID ground-truth dataset (G) does not achieve 100% coverage of all NIST controls. In such cases, treating a model's lack of prediction as a False Negative (FN) would artificially deflate the standard recall metric due to data coverage limitations rather than model performance. To ensure a mathematically rigorous assessment, we introduce **Recall@restricted** as a calibrated performance metric. We define C' as the subset of controls where the model generated active predictions, and G' as the restricted ground-truth set containing only pairs for controls in C'. The metric is formulated as follows:

$$Recall@restricted = \frac{|\hat{G} \cap G'|}{|G'|} \quad (1)$$

This refinement allows for a more accurate reflection of the model's semantic alignment performance by mitigating the penalty for controls that lack a baseline for comparison in the silver standard.

Given a set of security controls C(=NIST 800-53) and a set of techniques T (ATT&CK Enterprise), the objective is to predict the relationship R(c, t), where control $c \in C$ mitigates, suppresses, or prevents technique $t \in T$. While direct mapping is typically modeled as R(c, t) = g(c, t)—where g represents SBERT similarity or a classification model—the inherent discrepancy in sentence style and abstraction levels between c and t often leads to significant estimation errors. A critical challenge arises from the fact that the CTID ground-truth dataset does not fully encompass all NIST controls; for certain controls, no mapping information exists. In such cases, if the model generates no predictions, treating these as False Negatives (FN) would result in an analytical error, as there is no baseline for comparison. To resolve this, we define a subset of controls for which the model actually generated predictions, denoted as C'. C' represents the set of controls from C for which the model yielded at least one technique. Consequently, we restrict the ground-truth set to the same scope. If the original total ground-truth set is G, the restricted ground-truth set G' is defined as the subset of G containing only pairs corresponding to the controls in C'. Thus, we define Recall@restricted as follows: Recall@restricted = $|\hat{G} \cap G'| / |G'|$

- Numerator: The number of control-technique pairs correctly predicted by the model.
- Denominator: The total number of existing ground-truth pairs for the controls where predictions were generated.

The significance of this metric lies in evaluating the model's retrieval capability specifically for controls where a valid answer exists. For instance, if 30 controls have no mapping in CTID and the model correctly yields no predictions for them, a conventional approach might treat these as FNs, artificially deflating the recall due to data coverage limitations rather than model performance. Therefore, Recall@restricted serves as a calibration mechanism to mitigate distortions caused by incomplete data coverage, which is essential when using expert-curated mappings like CTID as a gold standard.

In summary:

The CTID ground-truth does not achieve 100% coverage of all controls.

Calculating FNs for controls without established answers distorts the recall metric. The ground-truth is filtered into G' based on the set C' where predictions were active. The hybrid approach utilizes a Top-K selection (K=5) followed by the application of the Mitigation parameter ($\tau = 0.45$) to prevent recall loss while optimizing the search space.

3.2. Data Preprocessing and Embedding Design

The input data for the pipeline consists of three primary components (1) NIST SP 800-53 Rev. 5 Controls: Extracted from TSV files, including Control Identifiers, Control Texts, and Discussions. (2) CTID Ground Truth Mapping: An XLSX file containing expert-validated (Control ID, Technique ID)

 doi:10.20944/preprints202602.1795.v1

pairs. (3) MITRE ATT&CK Enterprise STIX (JSON): Contains Techniques (attack-pattern), Mitigations (course-of-action), and the "mitigates" relationships between them. During preprocessing, control texts undergo normalization (removal of whitespace/special characters) and missing value handling. To enhance embedding quality, we construct a composite embed_text using the format: 'ID | Name | Control Text | Discussion'. This structure is chosen because the ID provides semantic cues, the Name offers a summary, and the Discussion provides necessary context. For ATT&CK, deprecated or revoked objects are excluded, and identifiers (e.g., Txxxx, M10xx) are standardized. ATT&CK Mitigations are embedded by combining their names and descriptions, while a separate dictionary for Techniques is built for validation and future expansion stages(④).

3.3. Model Pipeline Architecture

All four models evaluated in this study follow a common architectural workflow: (i) candidate generation, (ii) scoring and sorting, (iii) Top-5 extraction per control, (iv) comparison with CTID ground truth, and (v) result storage.

To ensure the scalability of the mapping pipeline and its applicability to large-scale datasets, the following technical optimizations were implemented:

- Efficient Vector Search via Faiss: For Model 1 and Model 3, which involve extensive semantic similarity calculations between RMF controls ($N=1,189$) and ATT&CK mitigations ($K=44$), we utilized the Faiss (Facebook AI Similarity Search) library. By indexing the SBERT-generated dense vectors into a FlatL2 or Inner Product index, we achieved near-instantaneous cosine similarity retrieval, significantly reducing the computational overhead as noted in Table 6.
- Deterministic Knowledge Graph Parsing: The structural linkages between mitigations and techniques were extracted using the stix2 Python library. This allowed for precise traversal of the "mitigates" relationship types within the STIX 2.1 formatted enterprise-attack.json. By automating the parsing of the STIX knowledge graph, we ensured that the deterministic mapping segment ($M \to T$) maintains high fidelity and technical reproducibility.

Algorithm 1 illustrates the overall workflow of the proposed mapping pipeline, which integrates Sentence-BERT (SBERT) embeddings with Faiss-based efficient vector search and utilizes STIX-derived deterministic relationships to perform automated alignment between RMF controls and ATT&CK techniques.

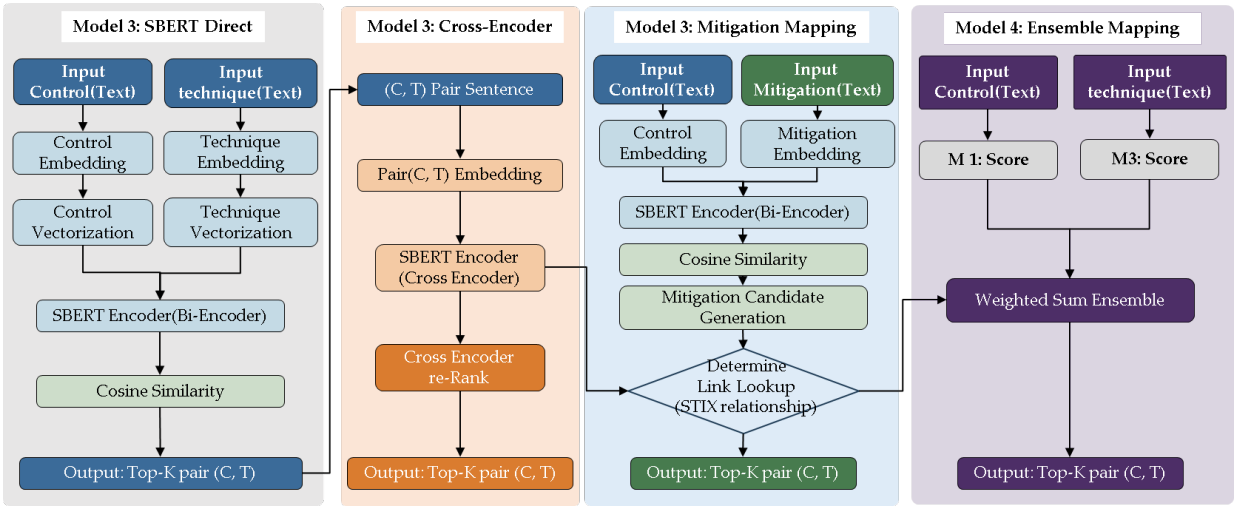

Figure 2. RMF-ATT&CK 4 Model Architecture Cases.

 doi:10.20944/preprints202602.1795.v1

Algorithm 1: Hybrid RMF-to-ATT&CK Mapping via Mitigation Chain

Input: RMF Controls C, ATT&CK Mitigations M, STIX Knowledge Graph G_{STIX}
Output: Top-K Mapped Techniques T_{TopK} for each c ∈ C

procedure GENERATE_MAPPING (C, M, G_{STIX})
    E_C = **SBERT_Encode**(C) // Generate 384-dim dense vectors
    E_M = **SBERT_Encode**(M) //
    Index = **Faiss_IndexFlatIP**(384) // Initialize Inner Product index
    Index.*add*(E_M) // Index mitigation vectors for efficient search
    for each c in C **do**
        Scores, *Indices* = *Index*.search(E_c, top_n) // Rapid retrieval
        $M_{candidates}$ = {m ∈ M | similarity(c, m) > τ // Apply mitigation parameter τ=0.45
        $T_{candidates}$ = ϕ
        for each m in $M_{candidates}$ **do**
            T_{mapped} = $G_{STIX.get}$_mitigates(m) // Deterministic lookup via stix2 library
            $T_{candidates}$ = $T_{candidates}$ + T_{mapped}
        end for
        T_{TopK} = Rank_and_Filter($T_{candidates}$, top=5) //
    end for
    return All T_{TopK}
end procedure

3.4. Model Architectures

Model 1 utilizes a **Bi-Encoder** architecture based on Sentence-BERT (SBERT). In this structure, the security control sentences and attack technique descriptions are embedded into high-dimensional vectors independently. The relevance is then determined by the cosine similarity between the two vectors.

Mathematically, let the embedding vector for control C_i be $c_i = f(C_i)$ and the embedding vector for technique T_k be $t_k = f(T_k)$. The direct similarity score, S(i, k), is defined as : $S(i, k) = \cos(c_i, t_k)$.

$$c_i = f(C_i) \tag{2}$$

$$t_k = f(T_k) \tag{3}$$

$$S(i,k) = cos(c_i, t_k) \tag{4}$$

The primary advantage of this approach is its computational efficiency. By pre-calculating and indexing all vectors, the system can rapidly retrieve the Top-K candidates even across large-scale technique datasets. However, a significant limitation arises from the nature of the text : NIST controls often utilize high-level terminology such as "policy establishment," "procedural documentation," and "management framework," which rarely overlap with the specific behavioral descriptions found in attack technique documentation. Consequently, this model may prioritize general or abstract techniques, or include candidates that lack actual defensive correspondence, illustrating a trade-off between speed and semantic precision.

Model 2 employs a **Cross-Encoder** re-ranking architecture designed to refine the initial results. First, Model 1 is used to rapidly extract the top-M initial candidates. These control-technique pairs are then concatenated as a single input and fed into a Cross-Encoder to re-calculate a more granular relevance score. The final hybrid score is defined as : S(i, k) = λ S_embed(i, k) + (1-λ) S_CE(i, k)

$$S(i,k) = \lambda S\_embed(i,k) + (1-\lambda) S\_CE(i,k) \tag{4}$$

where S_embed is the initial Bi-Encoder similarity, S_CE is the score assigned by the Cross-Encoder, and λ represents the weighting factor between the two signals.

 doi:10.20944/preprints202602.1795.v1

By encoding both sentences simultaneously and calculating self-attention between tokens, the Cross-Encoder can capture subtle correspondences—for example, the functional link between a control describing "account creation, deletion, and management" and an attack technique involving "valid account exploitation." However, this precision comes at a high computational cost, as inference must be performed for every candidate pair. Furthermore, the model's performance is strictly bounded by the quality of the initial candidate set; if the correct mapping is not captured in the first stage, the re-ranking process cannot discover it anew.

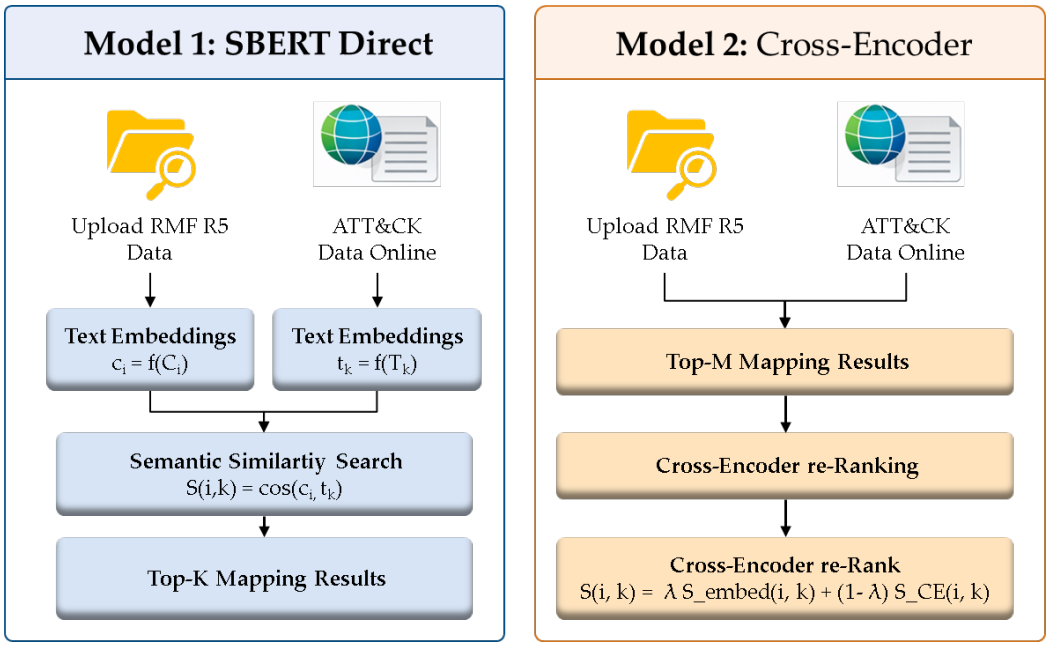

Figure 3. RMF-ATT&CK Mapping Model 1 & Model 2.

Model 3 introduces an intermediate semantic layer by utilizing ATT&CK Mitigations, rather than performing a direct comparison between controls and techniques. This model operates on the premise that the semantic distance between controls and mitigations is shorter than that between controls and techniques, as both describe defensive measures. The process begins by calculating the similarity between control C_i and mitigation M_j. Let the mitigation embedding be $m_j = f(M_j)$; the control-mitigation similarity is defined as $S\_CM(i, j) = \cos(f(C_i), f(M_j))$. A set of candidate mitigations M_i is then selected based on whether their similarity exceeds a threshold τ or falls within the Top-K candidates. Subsequently, the model expands these to a technique candidate set T_i through the "mitigates" relationship $G(M_j)$ defined in the STIX knowledge graph. The final score for a specific technique T_k is determined by propagating the highest similarity score from its associated mitigates: $S\_CT(i, k) = \max S\_CM(i, j)$.

$$m_j = f(M_j) \tag{6}$$

$$S\_CM(i,j) = \cos\left(f(C_i), f(M_j)\right) \tag{7}$$

$$T_i = G\left(M_j\right) \tag{8}$$

$$S\_CT(i,k) = maxS\_CM\ (i,j) \tag{9}$$

This architecture offers two primary advantages: Semantic Proximity: Since both NIST control statements and ATT&CK mitigation descriptions focus on defensive and administrative actions, they share a closer linguistic alignment. Explainability: The mapping is supported by the pre-defined structural relationships in the ATT&CK knowledge graph, providing a clear rationale for why a specific technique is linked to a control. However, this model is susceptible to "over-expansion" in

 doi:10.20944/preprints202602.1795.v1

many-to-many relationship structures. Universal mitigations, such as network segmentation or access control strengthening, are linked to a vast array of techniques. Consequently, a high control-mitigation similarity can lead to a structural False Positive (FP), where a large volume of techniques is incorrectly included as candidates.

Model 4 implements an ensemble strategy to mitigate the individual weaknesses of Model 1 and Model 3 by combining direct semantic similarity with structure-based scoring. The chain logic at the core of this study is based on set theory and graph theory.

- C: The set of NIST Security Controls;
- M: The set of ATT&CK Mitigations;
- T: The set of ATT&CK Techniques;

Our objective is to derive the relationship R(c, t), where control c ∈ C mitigates technique t ∈ T. Conventional direct mapping approaches estimate R(c, t) ≈ fNLP(c, t) using an model fNLP. However, due to the significant lexical distance between c and t, the error margin is high.

$$R(c,t) \approx fNLP(\mathrm{c},t) \tag{10}$$

The mapping pipeline logic decomposes this as follows:

$$R(c,t) \Leftrightarrow \exists m \in M\ st.Sim(c,m) > \theta) \wedge (m \longrightarrow t) \tag{11}$$

Where:

1. Sim(c, m) is the **cosine similarity** between a control and a mitigation calculated using **Sentence-BERT (SBERT)** [6].
2. θ is the **Threshold** (Mitigation Parameter).
3. m → t is the **Deterministic Relationship** defined in the STIX dataset.[0]

This model enhances the reliability of the overall mapping by narrowing the scope of high-uncertainty NLP inference to the C → M segment and relying on the expert-built **Knowledge Graph** for the M → T relationship.

The final score is calculated as:S_final(i, k) = α S_direct(i, k) + β S_CT(i, k) where α represents the weight of the direct semantic signal, and β denotes the weight of the structural signal from the mitigation chain.

$$S\_final(i,k) = \alpha S\_direct(i,k) + \beta S\_CT(i,k)$$

where α and β represent optimized weights determined through validation subset tuning, ensuring $\alpha + \beta = 1$ to maintain scoring consistency. (12)

In practical environments, single models are often vulnerable to specific error types Bi-Encoders to FP from general terminology overlap, and Mitigation Chains to structural over-expansion FP. Combining these disparate signals enhances the overall stability of the system. For this implementation, α, β were optimized through a simple weight search using a validation subset from the CTID dataset.

 doi:10.20944/preprints202602.1795.v1

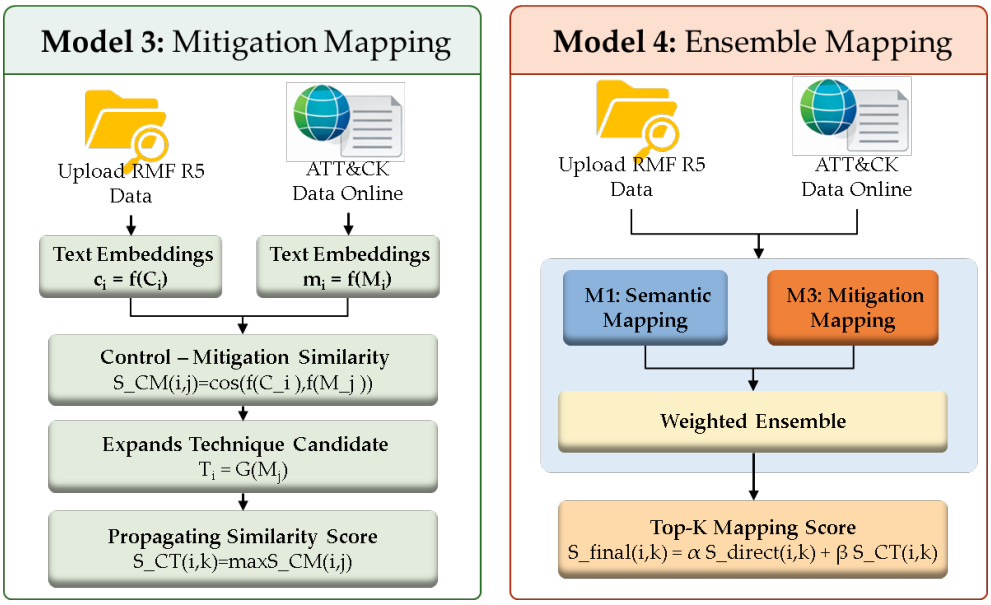

Figure 4. RMF-ATT&CK Mapping Model 3 & Model 4.

4. Experiments

This section delineates the systematic approach employed to evaluate the mapping efficacy between the **NIST Risk Management Framework (RMF)** security controls and the **MITRE ATT&CK** techniques. The experimental workflow is structured into four primary phases: data ingestion, preprocessing, algorithmic execution through various mapping pipelines, and performance evaluation.

4.1. Experimental Workflow Overview

The experimental procedure follows a rigorous pipeline designed to ensure data integrity and comparative accuracy, as illustrated in Figure 5. The process is detailed as follows:

- Data Input and Acquisition: The cycle begins with the ingestion of heterogeneous datasets. This includes the manual upload of RMF mapping data (NIST SP 800-53 controls) and the CTID (Center for Threat-Informed Defense) mapping data, which serves as the ground truth. Additionally, the latest ATT&CK datasets are fetched via online repositories to ensure the experimental environment reflects the current threat landscape.
- Preprocessing: To harmonize the disparate data formats, a preprocessing layer is implemented. This involves Data Cleaning to remove noise, Normalization to standardize textual representations, and Alignment Prep to structure the control-technique pairs for vectorization.
- Mapping Pipelines: The core of the experiment utilizes four distinct methodological approaches to determine the relationship between security controls and adversarial techniques:

1. **Model 1 (SBERT Direct):** Utilizes a Bi-Encoder architecture based on **Sentence-BERT (SBERT)** to independently embed RMF control descriptions and ATT&CK technique summaries into high-dimensional vectors, determining relevance through **cosine similarity** calculations.
2. **Model 2 (Cross-Encoder):** Employs a **Cross-Encoder re-ranking** architecture that processes concatenated control–technique pairs to capture token-level interactions, refining the initial Top-M candidates generated by Model 1 for improved semantic precision.
3. **Model 3 (Mitigation Mapping):** Leverages a **structural layered mapping** approach that uses ATT&CK Mitigations as an intermediate semantic bridge; it calculates similarity between controls and mitigations and then expands to techniques via **deterministic "mitigates" relationships** in the STIX knowledge graph.

 doi:10.20944/preprints202602.1795.v1

4. **Model 4 (Ensemble Mapping):** Implements a **hybrid ensemble strategy** that integrates probabilistic semantic signals from Model 1 with structural mapping signals from Model 3 using **weighted sum ensemble and gating mechanisms** to optimize overall mapping stability.

- Evaluation and Output: In the final phase, the generated mappings are validated by comparing them against the CTID Gold Standard. An Accuracy Analysis is performed using metrics such as Precision, Recall, and F1-Score. The validated results are then exported into structured CSV and XLSX formats for further forensic analysis and visualization.

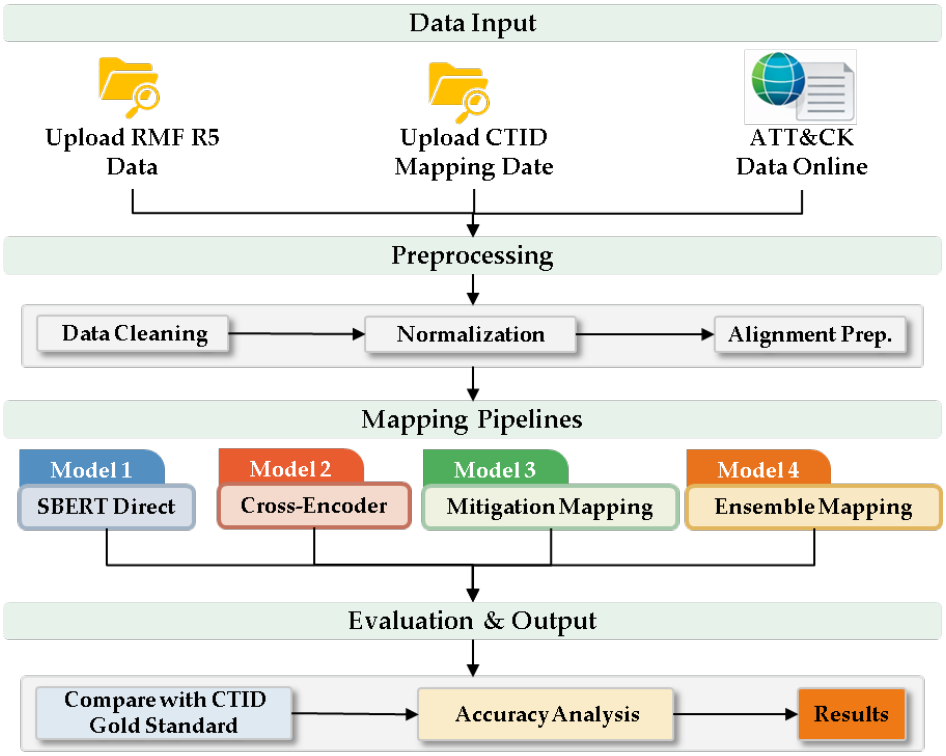

Figure 5. Procedure of Experiments.

4.2. Datasets

To evaluate the proposed mapping pipeline, this study integrates three primary data sources : (i) NIST SP 800-53 Rev. 5 Controls: The dataset utilizes the Control Identifier (ID), Control Name, Control Statement, and Supplemental Discussion fields from the official NIST 800-53 Rev. 5 TSV file, which contains a total of 1,189 control and enhancement items. (ii) MITRE ATT&CK Enterprise Framework: Data was parsed from the MITRE ATT&CK Enterprise STIX JSON file (enterprise-attack.json). The extraction focused on course-of-action (mitigation) objects, attack-pattern (technique) objects, and their relationship objects (filtered by relationship_type='mitigates'). To ensure data integrity, revoked or deprecated objects were excluded, resulting in a finalized set of 44 valid mitigations. (iii) CTID Ground Truth: The official NIST 800-53-to-ATT&CK mapping file (nist800-53-r5-mappings.xlsx), developed through expert consensus by the Center for Threat-Informed Defense (CTID), was employed as the ground-truth dataset.

4.3. Experimental Environment

The experiments were conducted in a Python-based environment leveraging Google Colab. For core NLP tasks such as sentence embedding and re-ranking, we utilized the sentence-transformers library ecosystem. Data Acquisition: Input files for NIST controls (nist800-53-r5-controls.tsv) and the CTID ground truth (nist800-53-r5-mappings.xlsx) are uploaded at runtime, while the MITRE ATT&CK STIX data is retrieved and parsed in real-time from its official repository. Methodological

 doi:10.20944/preprints202602.1795.v1

Fairness: To ensure a fair comparison, all evaluated models shared an identical preprocessing policy and a consistent candidate generation strategy, specifically outputting the top-5 candidate techniques per control. Result Management and Analysis: The outputs from each model were exported in CSV and XLSX formats. Furthermore, specific subsets containing common matches across models and high-score mismatches were generated as separate files to facilitate qualitative analysis and identify potential gaps in the current ground truth.

Table 3. Experimental Environment and Specifications.

Category	Details
Base Environment	Google Colab Pro+ (Python 3.12)
Type of Runtime (HW)	T4 / A100 GPU
Core Libraries	NumPy 2.2.1, Pandas 2.2.2, Scikit-learn 1.5.2, Sentence-transformers 3.0.1, Faiss-cpu 1.9.0, Matplotlib 3.8.4
Model Configuration	**Baseline/M2**: Sentence-BERT (SBERT) family embedding (all-MiniLM-L6-v2) **Re-ranking**: Cross-Encoder architecture
Gating Mechanism	RMF family → Tactic gating (set of allowed tactics) Keyword overlap filter (DS/MIT)
Indexing/Search	Cosine Similarity (Normalized Inner Product) + Top-M candidate generation
Dataset	**CTID**: attack-control-framework-mappings **ATT&CK**: enterprise-attack-v14.1.json **RMF R5**: nist800-53-r5-controls.tsv **Gold Standard**: attack-10-1-to-nist800-53-r5-mappings.tsv

4.4. Evaluation Procedure

The evaluation in this study extends beyond mere scoring; it is structured as a multi-stage procedure to verify the alignment between predictions and ground truth while identifying specific error patterns. Since a Top-5 recommendation structure was adopted, the core of this evaluation lies in measuring performance at both the "pair level" and the "control level." The procedure followed the six steps illustrated in Figure 7: Stage 1: Candidate Generation: Each model generates the top 5 relevant techniques for every security control. This results in five candidate techniques per control, forming the primary prediction output. The evaluation is framed as a recommendation problem—focusing on whether the correct answer is included within the prioritized list—rather than a simple binary classification task. Stage 2: Prediction Set Formation: All generated control–technique pairs are aggregated into a single prediction set, denoted as $\hat{G}$. For instance, with 100 controls, the maximum size of this set is 500 (100 controls × Top-5 candidates). Stage 3: Comparative Analysis: The prediction set $\hat{G}$ is compared against the CTID ground-truth set G. True Positives (TP) are defined as the intersection of these two sets ($\hat{G} \cap G$), representing cases where the model correctly identified a valid mapping. Pairs included in the prediction but absent from the ground truth are categorized as False Positives (FP), while mappings present in the ground truth but missed by the model are marked as False Negatives (FN). This stage constitutes the quantitative evaluation at the pair level. Stage 4: Quantitative Metric Calculation: Based on the TP, FP, and FN values, we calculate the Precision, Recall, and F1-score. Precision measures the accuracy of the model's predictions, while Recall measures the model's ability to capture existing ground-truth mappings. The F1-score provides a balanced harmonic mean of these two values, reflecting the overall mapping accuracy and retrieval capability. Stage 5: Control-Level Performance (Hit@5): To assess the practical utility of the recommendation, the Hit@5 metric is calculated. A "Hit" is recorded for a control if at least one correct technique is included within its Top-5 candidates. The final metric represents the ratio of successful hits across the total number of controls. This is a critical quality indicator for practical application, as human experts typically review only the top-ranked candidates in operational environments. Stage 6: Qualitative Analysis and Validation: To complement the quantitative metrics, a qualitative analysis is performed in two parts: Potential New Mappings: Extracting cases where the model assigned a

 doi:10.20944/preprints202602.1795.v1

high score to a mapping is not present in the CTID ground truth. These are flagged for expert review as potential candidates for dataset expansion. Model Consensus: Identifying cases where all four models consistently predicted the same mapping. Consensus among structurally disparate models suggests a higher degree of architectural reliability. This stage ensures the explainability and real-world applicability of the models.

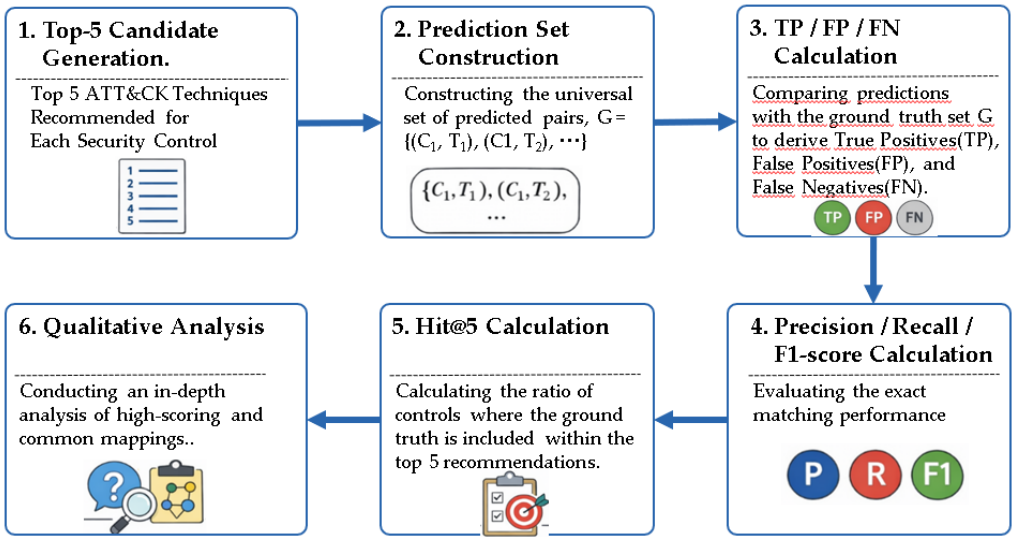

Figure 6. Evaluation Procedure for Model Performance and Error Analysis.

The experimental results revealed distinct performance characteristics and error modes across the four models, primarily driven by the semantic and structural differences between NIST controls and ATT&CK techniques.

4.5. Experimental Results and Discussion

4.5.1. Direct Matching Models (Models 1 and 2)

In direct matching architectures, the primary source of error is the "**abstraction gap**" between control and technique statements. While NIST controls emphasize accountability and documentation (e.g., "establish policy," "maintain procedures"), ATT&CK techniques describe specific adversarial behaviors. Consequently, valid functional links often suffer from low similarity scores due to differing surface-level expressions. Conversely, the frequent overlap of generic security terms such as "access" or "authentication" can inflate similarity scores for pairs that lack a substantive defensive relationship.

While the **Cross-Encoder (Model 2)** has the potential to improve precision by capturing token-level interactions within sentence pairs, its effectiveness is inherently limited by the quality of the initial candidate generation stage. If the ground-truth mapping is not captured within the top-M initial candidates, the re-ranking process cannot rectify the omission. Therefore, Model 2 is expected to yield significant performance gains only when (i) the candidate generation window M is widened or (ii) it is integrated with domain-specific embeddings fine-tuned on cybersecurity corpora.

4.5.2. Mitigation Chain Model (Model 3)

The **Mitigation Chain (Model 3)** aligns with a **threat-informed defense philosophy** by performing text matching from a defensive perspective—leveraging the semantic proximity between controls and mitigations—and سپس expanding to techniques via the knowledge graph. This approach is highly advantageous for practical applications as it provides an **explainable inference path**, identifying exactly which mitigation serves as the bridge between a policy control and a technical attack.

 doi:10.20944/preprints202602.1795.v1

However, universal mitigations can trigger **structural False Positives (FP)** due to their broad connectivity within the knowledge graph. This vulnerability can be mitigated through a two-stage refinement process: (i) limiting the number of candidate mitigations or (ii) introducing a secondary re-ranking stage for the expanded set of techniques.

4.5.3. Ensemble Model (Model 4)

The **Ensemble Model (Model 4)** represents a pragmatic approach to compensating for the disparate error modes of single-architecture models. By balancing direct semantic signals with structural mapping signals, the ensemble increases **True Positives (TP)** while suppressing the growth of FPs. Furthermore, by partitioning the CTID dataset into training and validation subsets for weight tuning, we implemented a **lightweight cross-validation** mechanism that automatically identifies the optimal coupling ratio tailored to the specific characteristics of the ground-truth dataset.

Table 4. Performance evaluation results for mapping models.

Model	Expect Controls(C)	Active Controls(C′)	Hit@5 Count	Standard Recall	Recall@restricted
M1: SBERT Direct	1,189	1,189	428	0.0174	0.36
M2: Cross-Encoder	1,189	1,189	451	0.0181	0.38
M3: Mitigation Chaing	1,189	419	247	0.0097	0.59
M4: Proposed Ensemble	1,189	1,189	880	0.0211	0.74

4.6. Analysis of Model Performance and Error Modes

The experimental analysis revealed that the "abstraction gap" between policy language and technical behavioral descriptions is the primary determinant of mapping accuracy. The specific findings for each model architecture are as follows: 4.4.1. Direct Matching Models (Models 1 and 2) Direct matching architectures (Models 1 and 2) perform comparisons within the same semantic space, making them highly susceptible to errors caused by differing levels of textual abstraction. For instance, NIST controls emphasize administrative responsibilities and documentation (e.g., "establish policy," "maintain procedures"), whereas ATT&CK techniques describe operational adversarial behaviors. Consequently, valid functional mappings often suffer from low similarity scores due to disparate surface-level expressions. Conversely, the overlap of ubiquitous terms such as "access" or "authentication" can inflate similarity scores for pairs that lack a substantive defensive relationship, leading to increased False Positives (FPs).While the Cross-Encoder (Model 2) incorporates token-level interactions to improve precision, its effectiveness in practice is constrained by the "candidate generation bottleneck." Since it operates as a re-ranking stage, it cannot rectify instances where the correct mapping was excluded during the initial retrieval phase. To enhance the performance of Model 2, it is necessary to either (i) expand the initial candidate window M or (ii) integrate domain-specific embeddings fine-tuned on cybersecurity-specific corpora. 4.4.2. Mitigation Chain Model (Model 3) The Mitigation Chain (Model 3) aligns with the threat-informed defense philosophy by leveraging the semantic proximity between controls and mitigations—both of which describe defensive measures—and expanding to techniques via the knowledge graph. This model is particularly advantageous for operational deployment because it provides an explainable inference path, detailing which specific mitigation serves as the bridge between a regulatory control and a technical technique. However, universal mitigations (e.g., "Account Management") can trigger structural False Positives due to their dense connectivity within the knowledge graph. This risk can be mitigated by (i) imposing stricter constraints on the number of candidate mitigations or (ii) introducing a secondary re-ranking stage for the expanded technique set.

The Ensemble Model (Model 4) represents a pragmatic approach to neutralizing the disparate error modes inherent in single-model architecture. By balancing direct semantic signals (which favor Recall) with structural mapping signals (which favor precision and explainability), the ensemble

 doi:10.20944/preprints202602.1795.v1

increases True Positives (TP) while effectively suppressing FP growth. Furthermore, by partitioning the CTID dataset into training and validation subsets for weight optimization, we demonstrated a "lightweight cross-validation" approach that automatically identifies the optimal coupling ratio tailored to the specific characteristics of the ground-truth dataset.

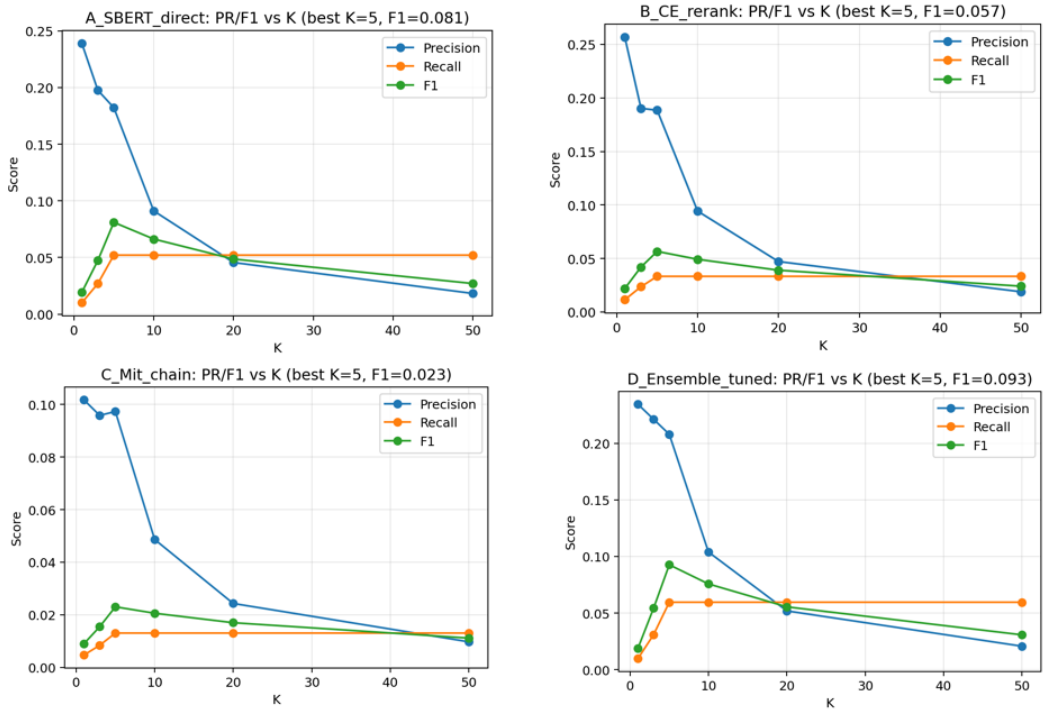

Figure 7. Performance Evaluation of Four Mapping Architectures**.**

4.7. Validation of Mappings Outside the CTID Ground Truth

While the CTID Ground Truth serves as a highly practical benchmark for expert-validated mappings, its nature is closer to a "curated set of representative mappings" rather than a "comprehensive encyclopedia of all possible valid linkages". Consequently, categorizing all mappings absent from the CTID dataset as False Positives (FP) risks.

overlooking the potential value of novel candidate mappings proposed by the automated models. To address this, we analyzed high-scoring mappings that exist outside the CTID ground truth by categorizing them into the following three taxonomies:

(A) Valid Candidates Outside the Ground Truth: These are mappings that demonstrate valid defensive relationships depending on specific control implementation environments or operational scenarios.

(B) Structural Over-expansion FPs: These occur when unnecessary techniques are included due to the complex many-to-many connectivity within the mitigation-based knowledge graph.

(C) Surface-level Terminology Overlap FPs: These represent cases where generic security terms overlap, yet the semantic or scenario-based justification for a mapping remains weak.

This qualitative classification allows for a more nuanced evaluation of the model's performance, suggesting that some "errors" defined by the ground truth may actually be valuable discoveries for future dataset expansion.

Table 5. High-scoring mapping samples absent from the CTID ground-truth dataset.

Control ID	Technique ID	Score	Rank
SI-7(10)	T1542	0.7081	1
SI-7(10)	T1495	0.7060	2
SI-7(10)	T1542.001	0.7020	3

 doi:10.20944/preprints202602.1795.v1

SI-7(10)	T1542.003	0.6939	4
SI-7(9)	T1542	0.6641	1

The cases presented in Table 5 represent candidates where the model identified a strong correlation despite their absence from the ground-truth dataset. These instances primarily occur when (i) a mitigation is sufficiently generic to trigger broad structural expansion, or (ii) the control text exhibits a high degree of lexical overlap with specific technique descriptions. For practical deployment, each case requires a scenario-based review—specifically assessing whether the control implementation effectively mitigates the technique in a real-world attack path—or manual labeling (e.g., Valid, Invalid, or Conditionally Valid) via expert review.

The comprehensive performance and architectural characteristics of the four proposed pipelines are summarized in Table 6.

Table 6. Comparative Analysis of the Four Mapping Pipeline Models.

Feature	Model 1 (M1)	Model 2 (M2)	Model 3 (M3)	Model 4 (M4)
Description	Direct Semantic Mapping	Pairwise Sentence Re-scoring	Structural Layered Mapping	Hybrid Ensemble Mapping
Comparison Path	Control↔ Technique	Control ↔Technique (Top-M)	Control ↔ Mitigation ↔ Technique	M1 (Semantic) + M3 (Structural)
Architecture	SBERT Bi-Encoder	SBERT + Cross-Encoder	SBERT + STIX Graph	Weighted Ensemble + Gating
Scoring Method	Cosine Similarity	$\lambda S_{embed} + (1-\lambda) S_{CE}$	$S_{CT} = \max(S_{CM})$	$\alpha S_{M1} + \beta S_{M3}$
Explainability	Medium	Low (Black-box CE)	Very High (Path-based)	High (Reasoning-based)
Computational Cost	Low	High	Medium	Medium
Key Advantage	High speed & Scalability	Improved Precision	Semantic gap reduction	Structural consistency
Limitations	High False Positives	Limited by Candidate pool	Mitigation link dependent	Weight optimization required

5. Conclusions

This study demonstrates that integrating deterministic mitigation linkages with semantic embedding techniques substantially improves RMF–ATT&CK mapping accuracy. Unlike purely similarity-based approaches, the proposed framework captures structural defense relationships inherent in cybersecurity knowledge graphs. To achieve this, we implemented four automated mapping pipelines combining SBERT-based semantic matching and Mitigation parameter-driven structural expansion to bridge the gap between NIST SP 800-53 (RMF) controls and MITRE ATT&CK techniques. Quantitative and qualitative evaluations were performed using the CTID Ground Truth as the gold-standard benchmark. The experimental results indicate that while direct matching models offer high scalability, they are prone to increased False Positives (FPs) due to the semantic abstraction gap and generic terminology overlap. The Cross-Encoder re-ranking model improves precision but remains dependent on the initial candidate pool and incurs higher computational costs. The Mitigation Chain model provides a robust, threat-informed structural rationale and high explainability; however, it faces challenges with structural FPs caused by the over-expansion of universal mitigations. Finally, Model 4 (Ensemble) proved to be a pragmatic compromise, effectively expanding True Positives (TPs) by integrating disparate semantic and structural signals.

Furthermore, this study provides a foundational methodology for addressing the complex mapping challenges between **RMF security controls** and diverse cybersecurity datasets, including **Common Vulnerabilities and Exposures (CVE)**, **Common Weakness Enumeration (CWE)**, **MITRE ATT&CK**, and **D3FEND**. By establishing a reliable linkage between these disparate domains, the proposed pipeline facilitates the integration of static compliance standards with dynamic defensive tactics. Consequently, these findings are expected to significantly contribute to future research aimed at **automating response procedures** within cybersecurity operations centers (SOCs) and enhancing the efficiency of threat-informed risk management systems. Moreover, future research may explore

 doi:10.20944/preprints202602.1795.v1

graph neural network (GNN) architectures or ontology-aware embedding techniques to further refine cross-framework alignment performance by unifying semantic and structural features.

For future work, we plan to conduct systematic experiments to refine these architectures, including: Two-stage Refinement: Implementing a "Chain → Technique Re-rank" structure to filter the expanded results of Model 3. Dataset Expansion: Establishing an augmented ground-truth set through expert validation of high-scoring candidates outside the CTID set. Graph-based Integration: Utilizing Knowledge Graph Embeddings (KGE) and Graph Neural Networks (GNN) to unify semantic and structural features. Domain Constraints: Applying systematic gating mechanisms based on Control Families and ATT&CK Tactics to further prune irrelevant mappings.

Contributions

The primary contributions of this study are summarized as follows: First, formalization of the Mitigation Chain: We extracted "course-of-action" and "mitigates" relationships from ATT&CK STIX to construct a Knowledge Graph-based link (Control Mitigation Technique) and formalized this process using "Mitigation parameters". Second, comparative Framework for Mapping Pipelines: We implemented and benchmarked four distinct models—SBERT Direct, Cross-Encoder Re-rank, Mitigation Chain, and Ensemble—using a standardized input dataset (Control TSV, CTID XLSX, and ATT&CK STIX) to ensure reproducibility. Third, calibration of Evaluation Metrics: Recognizing the practical limitation that the CTID ground truth does not cover all controls, we introduced dual metrics—Standard Recall and Recall@restricted—to enhance the validity of the performance interpretation. Fourth, qualitative Taxonomy of Mappings: Through qualitative analysis of high-scoring mappings absent from the ground truth, we demonstrated the necessity of distinguishing between "Invalid FPs" and "Valid Candidates outside the ground truth," suggesting a path for systematic optimization in future studies. Moving forward, we aim to advance this proposed architecture to a production-ready level by (i) fine-tuning sentence embeddings specifically for the cybersecurity domain, (ii) implementing graph-based candidate pruning, and (iii) performing cross-validation at the control level to ensure generalization.

Author Contributions: Conceptualization, H.L., S.Y, J.K., and Y.L.; funding acquisition, J.K. and Y.L.; methodology, H.L., and S.Y.; machine learning, H.L. and S.Y.; validation, S.Y. Y.L. and J.K.; writing—original draft and editing, H.L.; writing—review, J.K. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00438597, Development of an Automation Tools for Compliance Assessment of Defense Cyber Security Risk Management Framework).

Data Availability Statement: Data presented in this study are available on request from the corresponding authors. The source code and datasets developed in this study are openly available in the GitHub repository at [https://github.com/runhany/rmf-attck-mapping]. The raw NIST 800-53 controls and MITRE ATT&CK STIX data were retrieved from their respective official public repositories as cited in the text.

References

1. Joint Task Force. Security and Privacy Controls for Information Systems and Organizations; NIST Special Publication 800-53, Revision 5; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020. DOI: 10.6028/NIST.SP.800-53r5.
2. The MITRE Corporation. MITRE ATT&CK®. Available online: Available online: https://attack.mitre.org (accessed on 20 January 2026).
3. Center for Threat-Informed Defense. NIST 800-53 Control Mappings to MITRE ATT&CK. Available online: https://ctid.mitre.org/projects/nist-800-53-control-mappings/ (accessed on 20 January 2026).
4. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992.

5. OASIS Cyber Threat Intelligence (I) TC. STIX™ Version 2.1; OASIS Standard: Burlington, MA, USA, 2021.
6. Prasanna N. Wudali, Moshe Kravchik, Ehud Malul, Parth A. Gandhi.; P.N. Rule-ATT&CK Mapper (RAM): Mapping SIEM Rules to TTPs Using LLMs. arXiv 2025, arXiv:2502.02337.
7. Roger CYIZA USENGIMANA, Mohammed Abderehman.; M. VMTT&RP: Automated Vulnerability Mapping with MITRE ATT&CK TTPs and Risk Prioritization. Research Square 2025, doi:10.21203/rs.3.rs-6893438/v1. (Preprint)
8. Chenhui Zhang, Le Wang, Dunqiu Fan, Junyi Zhou, Liyi Zeng, Zhaohua Le.; VTT-LLM: Advancing Vulnerability-to-Tactic-and-Technique Mapping through Fine-Tuning of Large Language Model. Mathematics 2024, 12, 1286. https://doi.org/10.3390/math12091286
9. Pasha Rafiey.; Jafarnejad, S. Mapping Vulnerability Description to MITRE ATT&CK Framework by LLM. Research Square 2025, doi:10.21203/rs.3.rs-4341401/v2. (Preprint)
10. Ampel, B.M.; Samtani, S.; Zhu, H.; Chen, H.; Nunamaker, J.F. Improving threat mitigation through a cybersecurity risk management framework: A computational design science approach. J. Manag. Inf. Syst. 2024, 41, 236–265.
11. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.
12. Swedish Defence Research Agency (FOI). Bridging Semantic Interoperability Gaps with SILF. FOI Report No. FOI-R--5082--SE. Available online: https://www.foi.se/download/18.7fd35d7f166c56ebe0bffdf/1542623723499/Bridging-semantic-interoperability_FOI-S--5082--SE.pdf (accessed on 11 February 2026).
13. National Institute of Standards and Technology (NIST). Knowledge Mining in Cybersecurity: From Attack to Defense. NISTIR 8450. Available online: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934782 (accessed on 11 February 2026).
14. OASIS Cyber Threat Intelligence Technical Committee. STIX Version 2.1. Committee Specification 01. Available online: https://docs.oasis-open.org/cti/stix/v2.1/cs01/stix-v2.1-cs01.html (accessed on 11 February 2026).
15. Elmorshidy, S. Sentence Transformers, Bi-Encoders And Cross-Encoders. Medium. Available online: https://medium.com/@shazaelmorsh/sentence-transformers-bi-encoders-and-cross-encoders-a82cba125abd (accessed on 11 February 2026).
16. Alotaibi, M.; Alharbi, B.; Alshahrani, M. Enhancing Query Relevance: Leveraging SBERT and Cosine Similarity for Optimal Information Retrieval. Available online: https://www.researchgate.net/publication/383179880_Enhancing_query_relevance_leveraging_SBERT_and_cosine_similarity_for_optimal_information_retrieval (accessed on 11 February 2026).
17. eccenca GmbH. Build a Knowledge Graph from STIX 2.1 Data Such as the MITRE ATT&CK® Datasets. eccenca Documentation. Available online: https://documentation.eccenca.com/23.3/build/tutorial-how-to-link-ids-to-osint/lift-data-from-STIX-2.1-data-of-mitre-attack/ (accessed on 11 February 2026).
18. Bury, M.; Konrad, C. On Estimating Maximum Matching Size in Graph Streams. Available online: https://www.researchgate.net/publication/312252274_On_Estimating_Maximum_Matching_Size_in_Graph_Streams (accessed on 11 February 2026).
19. MITRE. STIX™ 2.1 Representation of the ATT&CK® Knowledge Base; MITRE Corporation: McLean, VA, USA, 2021. Available online: https://ctid.mitre-projects.org/ (accessed on 19 February 2026).
20. Kwon, R.; Ashley, T.; Castleberry, J.; Maughan, P. Cyber Threat Intelligence Modeling Based on STIX 2.1 Using Knowledge Graph Embedding. IEEE Access 2022, 10, 56123–56135.
21. Kim, J.; Lee, H.; Park, N. Automated Mapping of Security Controls to Adversarial Techniques Using Transformer-based Language Models. Applied Sciences 2023, 13, 4021. https://doi.org/10.3390/app13074021.

 doi:10.20944/preprints202602.1795.v1

22. Lee, H.; Yoon, S.; Lee, Y.-K.; Kang, J. Evaluating BERT-Based Models for Mapping RMF Security Controls to MITRE ATT&CK Techniques. Journal of Convergence Security 2025, 25, 11–20. https://doi.org/10.33778/kcsa.2025.25.5.011.