AI System For MITRE ATT&CK Threat Classification

Presented by
Prapatsorn Alongkornpradub - st124846
Vorameth Reantongcome - st124903
Ekkarat Techanawakarnkul - st124945

Presented on 28 April 2025

AT 82.05 AI: Natural Language Understanding
Asian Institute of Technology

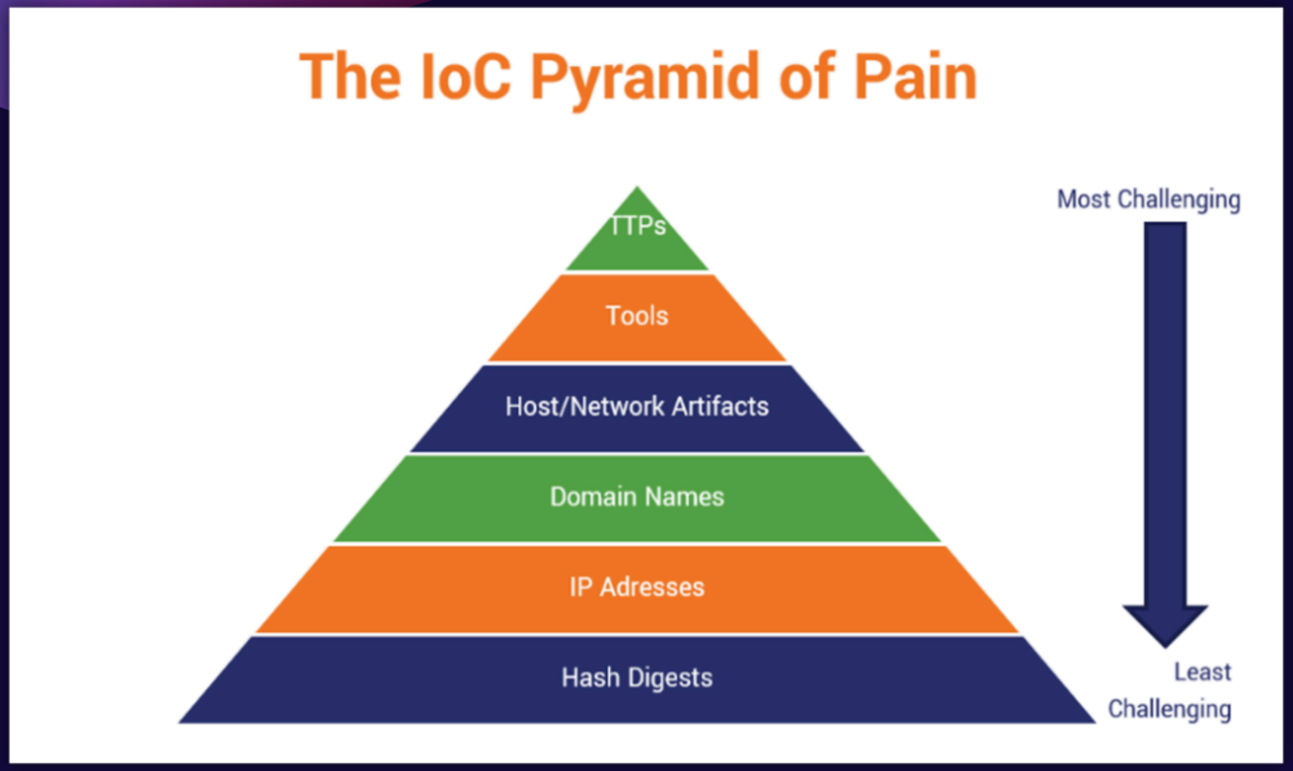
Agenda

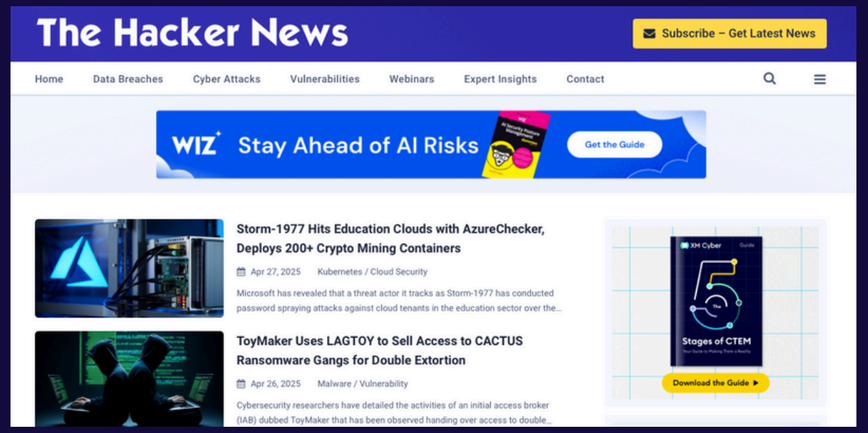
1. Introduction
2. Research Questions
3. Related Works
4. Methodology
5. Results
6. Discussion
7. Conclusion
8. Limitations and Challenges
9. Future Work

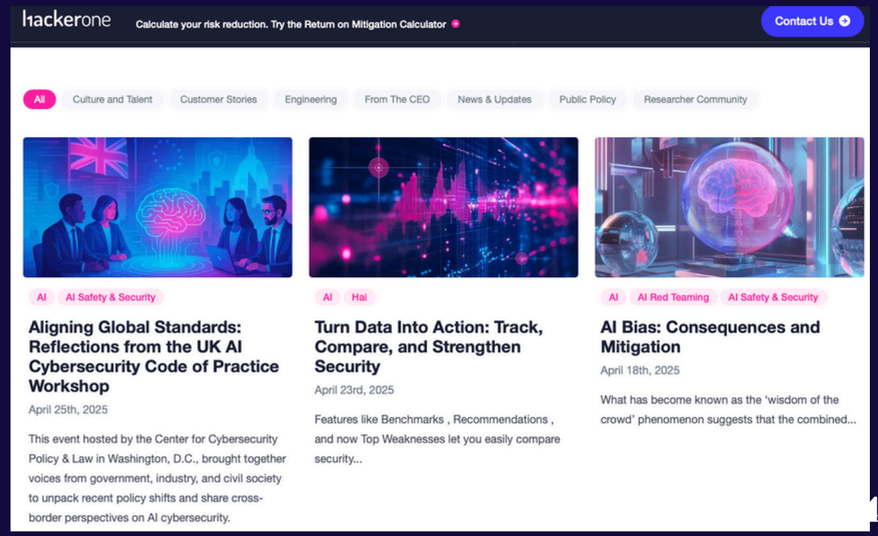
Introduction

In order to fasten the process of cybersecurity expert to identify the modus operandi of the attackers through cyber-intelligent news, with a assistant from the Natural Language Processing could support and enhance process of attack identification allowing work to be distributively arranged into focusing mitigate the risk, identify impact, and other consequence procedures.

How to detect ?



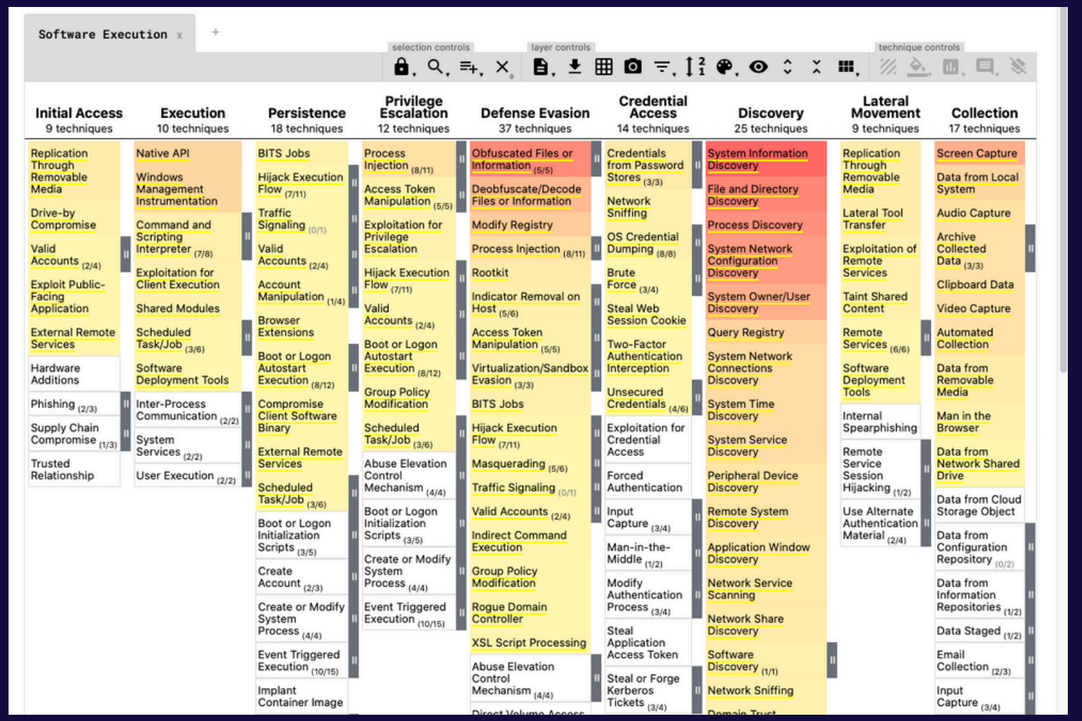




Source : Indicators of Compromise: Cybersecurity's Digital Breadcrumbs

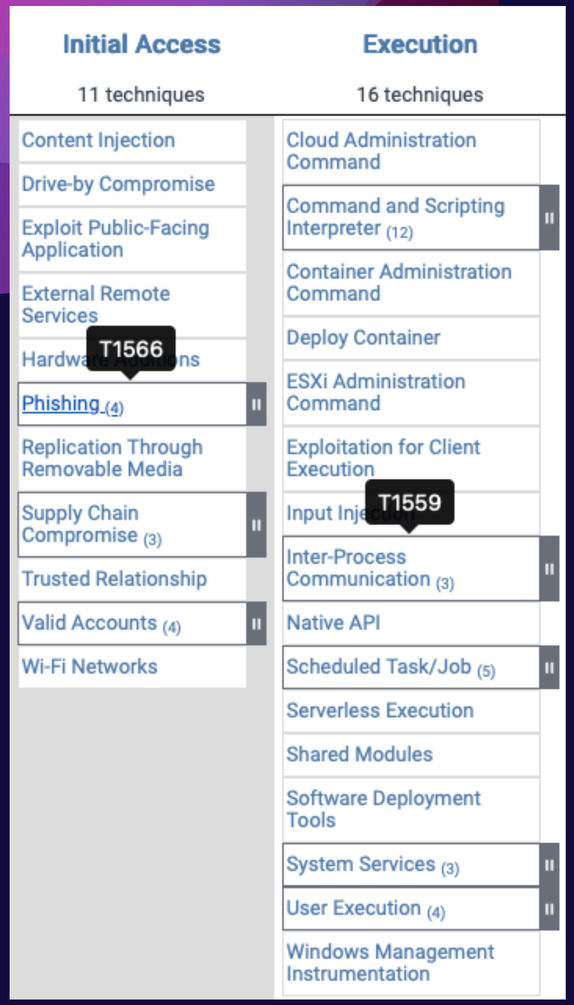
What is TTPs ?

- TTPs stands for Tactics, Techniques, and Procedures
- To centralized and accumulate all the TTPs in the cyber world, organization called MITRE provides a ATT&CK framework which is the knowledge base for describing any attack behaviour.



Source : MITRE ATT&CK

MITRE ATT&CK



Inter-Process Communication

Sub-techniques (3)

ID	Name
T1559.001	Component Object Model
T1559.002	Dynamic Data Exchange
T1559.003	XPC Services

Source : Indicators of Compromise: Cybersecurity's Digital Breadcrumbs

Research Questions

1. How does the system's classification accuracy compare to the existing chatbots?

2. How well can the AI model understand and match complex attack descriptions in cybersecurity news articles to the MITRE ATT&CK techniques, even when clear keywords aren't used?

Related Works

Key Points:

- Pre-trained models like **CTI-BERT** have revolutionized threat classification but face limitations in dataset and language diversity. (Model)
- Frameworks like **NCE** and deep learning methods are essential in addressing multi-label challenges and improving classification performance. (Datasets)
- Datasets such as **AnnoCTR** and **CySecED** are invaluable resources for training models to classify cybersecurity threats effectively. (Datasets)
- **NER** and attention-based mechanisms, like **quasi-attention**, are enhancing model accuracy in cybersecurity tasks by capturing contextual information. (Model)

Related Works

1. Pre-trained Models and Classification Frameworks

- **CTI-BERT:** Developed for cyber threat intelligence, outperforms general models in MITRE ATT&CK classification (Park and You, 2023).
- **Limitations:** Dataset size and language coverage.
- **Full-stack NLP Pipeline:** Extracts threat intelligence from unstructured texts but lacks MITRE ATT&CK integration (Park and Lee, 2022).

2. Multi-Label Challenges in MITRE ATT&CK Classification

- **NCE Framework:** Introduces a Noise Contrastive Estimation approach to handle multi-label challenges and missing annotations, showing improved performance (Nguyen et al., 2024).
- **Deep Learning vs. Traditional Approaches:** Deep learning methods outperform traditional methods at the sentence level (Orbinato et al., 2022).

Related Works

3. Cybersecurity Datasets and Entity-Level Extraction

- **AnnoCTR Dataset:** Annotated dataset for threat reports linked to MITRE ATT&CK, beneficial for training classification models (Lange et al., 2024).
- **CySecED:** Expands on earlier datasets, adding more event types and broader document-level context (Trong et al., 2020).
- **NER:** Named Entity Recognition improves classification by linking identified entities to MITRE ATT&CK taxonomies (Park and Lee, 2022).

4. Aspect-Based and Context-Aware Modeling

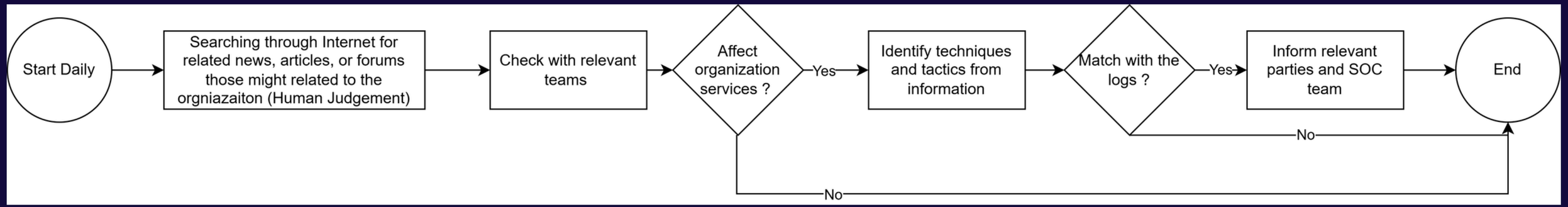
- **Attention Mechanisms:** Adjust attention to focus on relevant parts of a sentence for tasks like aspect-based sentiment analysis (Wu and Ong, 2020).
- **Quasi-Attention Mechanism:** A mechanism that allows the model to assign both positive and negative attention to words, helping reduce irrelevant influence while highlighting important information.
- **Use in Cybersecurity:** Improves understanding of the relationship between parts of a threat report and MITRE ATT&CK techniques.

Methodology

Old Workflow

1. Search and Filter:
- Search websites, articles, and forums for relevant cybersecurity incidents.
2. Team Review:
- Share the information with teams to assess potential impact.
3. Information Verification:
- Verify the relevance and accuracy of the information.

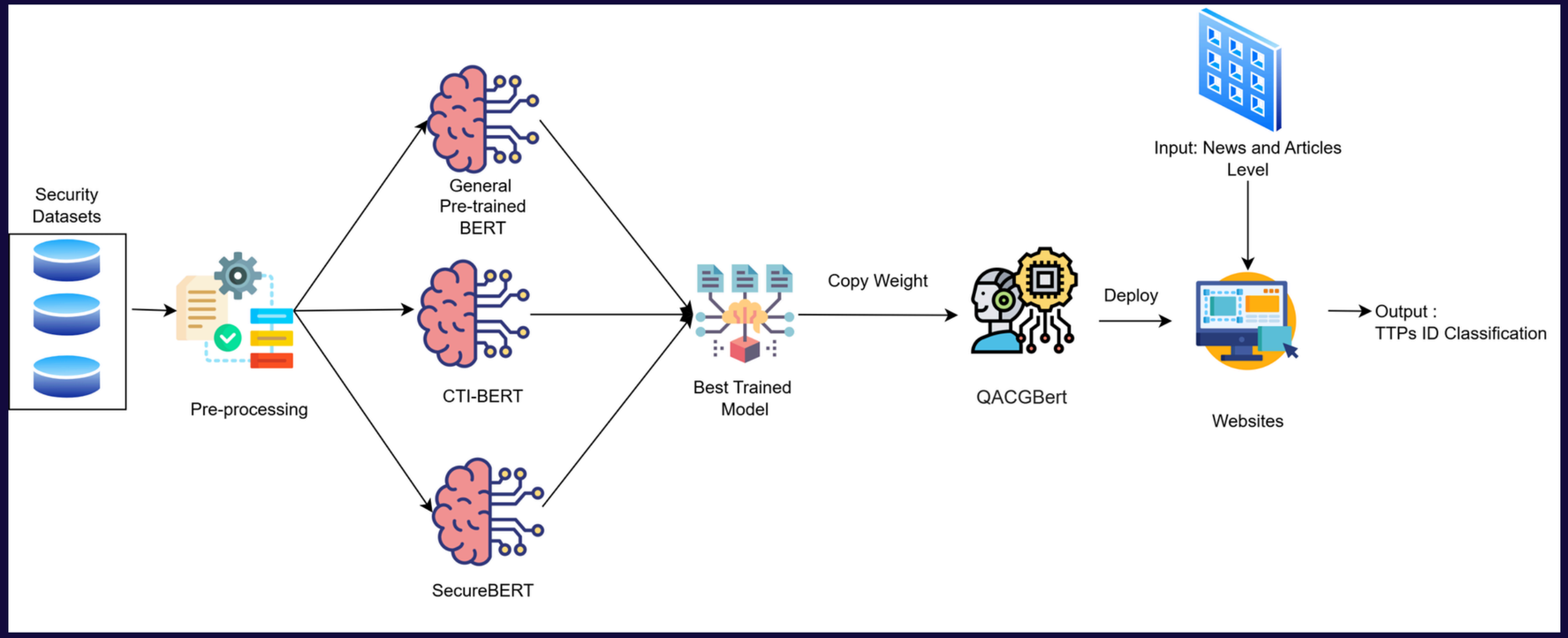
4. Extract Techniques and Tactics:
- Manually extract MITRE ATT&CK techniques and tactics from the incident details.
5. Pattern Matching:
- Compare the extracted techniques with known attack patterns to identify and classify the attack.
6. Alert Relevant Parties:
- Notify relevant teams, such as SOC, to initiate a mitigation plan.



Methodology

New Workflow

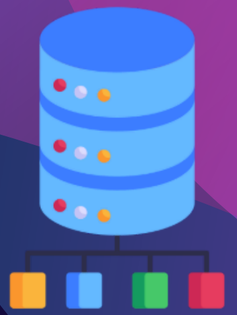
- New Workflow Improves the Old Workflow in Identification of Techniques and Tactics part.
- The old method required manual extraction and mapping of techniques and tactics. With the new workflow, the model classifies the content into MITRE ATT&CK techniques automatically threat identification.



Methodology

Datasets

There is only one dataset that will be used for training the model and using it as a document retrieval.



Tumeteor Dataset (tumeteor/Security-TTP-Mapping)

- Maps security-related text to Tactics, Techniques, and Procedures (TTPs).
- Helps identify how hackers might use similar patterns in real-world news.
- This dataset consists of 14.9K rows for train, 2.63K rows for validation, and 3.17K for test

Example of Datasets

text1 string · *lengths* 7↔257 84.8%	labels string · *lengths* 9↔19 94.9%
The command processing function starts by substituting the main module name and path in the hosting process PEB, with the one of the default internet browser. The path of the main browser of the workstation is obtained by reading the registry value	['T1057']
Along the way, HermeticWiper's more mundane operations provide us with further IOCs to monitor for. These include the momentary…	['T1569.002']
These Microsoft Office templates are hosted on a command and control server and the downloaded link is embedded in the first stage…	['T1584.004']

Methodology

Model

The following BERT-based models will be evaluated for their ability to classify cybersecurity data and identify MITRE ATT&CK techniques:

BERT-Base-Uncased

- Base BERT model available on Huggingface.
- Serves as the baseline for comparison.
- Will be evaluated against CTI-BERT and Secure-BERT.

CTI-BERT

- Specialized in the cybersecurity domain.
- Built from the SecBert model to focus on cyber threat intelligence.
- Primary model for the project.

Secure-BERT

- Extends from the pre-trained Ro-Bert model.
- Aimed at improving cybersecurity-specific tasks.
- Will be compared with other models for performance.

Methodology

Model

Context-Guided Quasi-Attention (QACG-BERT) model, focusing on how it enhances the standard self-attention mechanism in BERT for Targeted Aspect-Based Sentiment Analysis (T-(A)BSA)

The equations and concepts

1. Standard Self-Attention in BERT

$$\mathbf{A}^h_{\text{Self-Attn}} = \text{softmax}\left(\frac{\mathbf{Q}^h \mathbf{K}^{h^T}}{\sqrt{d_h}}\right)$$

Calculates the attention weights from the query and key.

2. Quasi-Attention (Context-Based)

$$\mathbf{A}^h_{\textit{Quasi}\text{-Attn}} = \alpha \cdot \text{sigmoid}\left(\frac{f_\psi(\mathbf{C}^h_Q, \mathbf{C}^h_K)}{\sqrt{d_h}}\right)$$

Quasi-Attention comes to improve the regular attention through context-based vectors that help the model understand relationships between words in the given context

3. Combining Regular Attention and Quasi-Attention

$$\hat{\mathbf{A}}^h = \mathbf{A}^h_{\text{Self-Attn}} + \lambda^h_A \mathbf{A}^h_{\textit{Quasi}\text{-Attn}}$$

Combine the standard attention and the quasi-attention using a **compositional factor λh** to control how much context affects the final attention

4. The Gating Factor

$$\lambda^h_A = 1 - (\beta \cdot \lambda^h_Q + \gamma \cdot \lambda^h_K)$$

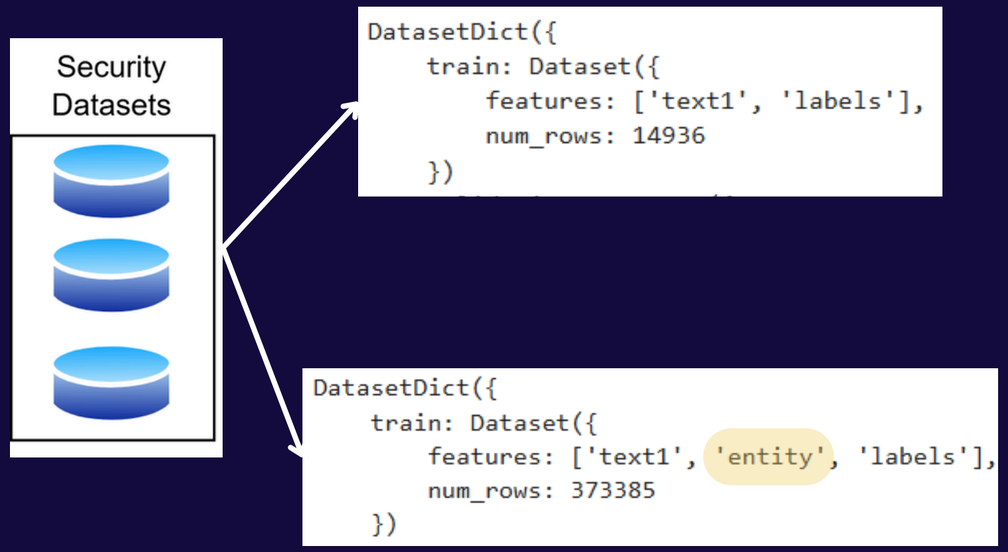
- β and γ are scalars that adjust the weight of the context-based attention
- λQ and λK are the bidirectional components that determine how much positive or negative impact the context has.

Methodology

Overview Experiment

From new workflow, there are two experiments including Input Pre-processing Experiment and Model Experiment

Input Pre-processing Experiment



First dataset, Sentence-Level Processing:
Each sentence in the dataset will be treated as a separate input for classification. This method is applied during the initial training of all baseline models (BERT-Base, CTI-BERT, Secure-BERT).

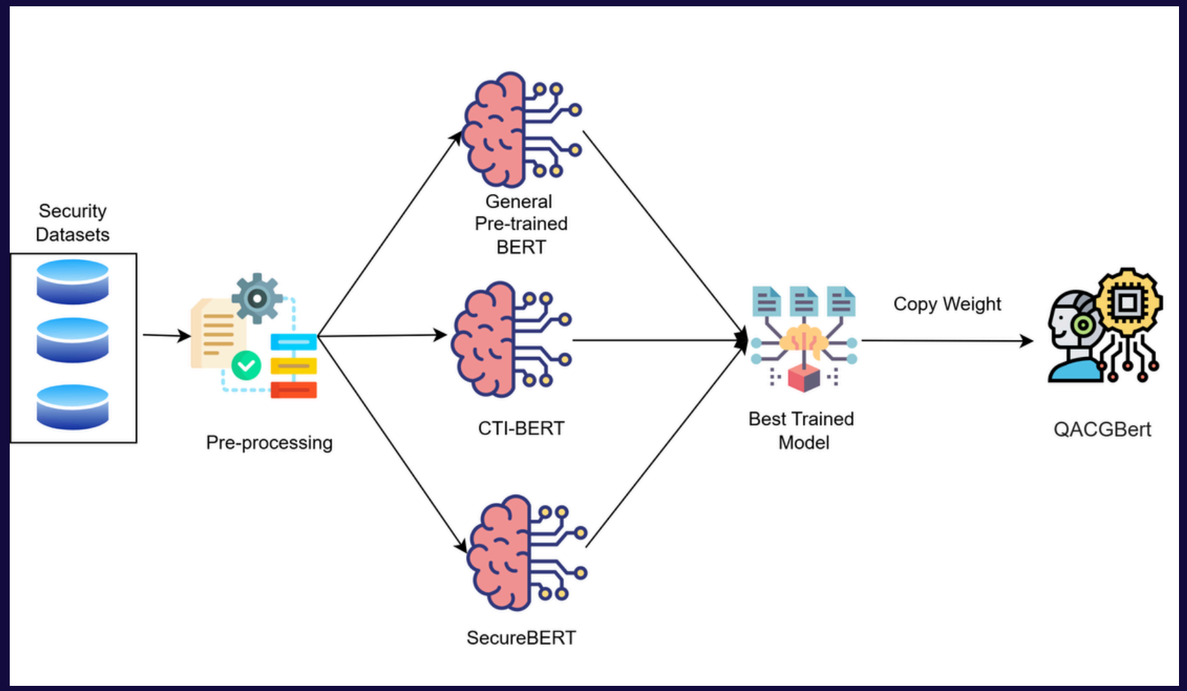
Second dataset, Entity-Augmented Processing:
Sentences are enriched with entity information extracted via Named Entity Recognition (NER) to highlight key cybersecurity terms. This method is used with the QACGBertForSequenceClassification model to enhance contextual understanding.

Methodology

Overview Experiment

From new workflow, there are two experiments including Input Pre-processing Experiment and Model Experiment

Model Experiment



1. Select Models for Comparison
- Three models selected for comparison are: BERT-Base-Uncased, CTI-BERT and Secure-BERT
2. Tokenizer and Word Embeddings
3. Training the Models: train 3 models with the same dataset and turn hyperparameters
4. Evaluate Performance
- Training loss, validation loss, and classification accuracy
5. Refine with QACGBertForSequenceClassification
- Selecting the best model to enhance performance, incorporating additional context information.

Methodology

New Workflow

Evaluation / Metrics

To evaluate the models, two evaluation methods are applied:.

Performance Metrics (Automated Evaluation)	• **Accuracy** – Measures overall correctness • **Precision** – Checks how many classified threats are relevant • **Recall** – Ensures the model identifies all actual threats • **F1-Score** – Balances precision & recall for overall performance
Top-K Ranking (NDCG Score)	• Normalized Discounted Cumulative Gain (NDCG) • Measures the accuracy of top-k recommended techniques, reflecting how well the model ranks the most relevant MITRE ATT&CK techniques.

Results

Table: **Standard Classification Metrics** Results of BERT-based Models

Metrics	BERT-base-uncased	CTI-BERT	SecureBERT	QACGBERT
Training loss	3.575500	2.774000	3.579800	1.706076
Validation loss	3.438241	2.707260	3.307534	4.723936
Accuracy	0.437643	0.546388	0.444487	0.168077
Precision	0.292122	0.431662	0.299807	0.163138
Recall	0.437643	0.546388	0.444487	0.168077
F1-Score	0.333476	0.466138	0.337116	0.144451

Results

Table: **Normalized Discounted Cumulative Gain (NDCG) Metric** Results of BERT-based Models

In order to calculate the accuracy score, the ground truth which created by using ChatGPT classification of TTPs will be compared on both CTI-BERT and QACGBERT probability score. The top k that will be used in this evaluation is 20.

Metrics	CTI-BERT	QACGBERT
NDCG Score wit k = 20	0.0000	0.0163

Discussion

Addressing Core Research Questions

RQ1: How does the classification accuracy of the proposed system compare to existing chatbots?

- **Findings:** The system's classification accuracy is currently lower than existing chatbots.
- **Challenges:** Insufficient training data and model generalization across different attack descriptions.
- **Next Steps:** Refine training processes and expand the dataset for better performance.

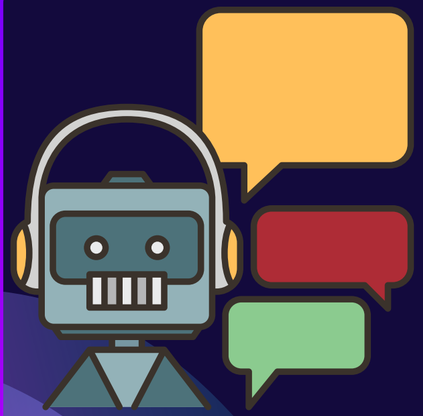
RQ2: How effectively can the AI model understand and match complex attack descriptions to MITRE ATT&CK techniques, particularly when clear keywords are not present?

- **Findings:** The model shows potential but struggles with complex and unclear attack descriptions.
- **Challenges:** Difficulty in generalizing to new or complex attack scenarios.
- **Next Steps:** Fine-tune the BERT model with a more diverse dataset.

Discussion

Impact of the Work

- **Current State:** AI system processes data faster but lacks consistent accuracy.
- **Goal:** Improve accuracy and assist security teams respond more effectively.
- **Web Application:** Allows security analysts to input unstructured text (cybersecurity news articles) and receive MITRE ATT&CK predictions.



Conclusion

The primary goal of this research was to **reduce the manual effort** required for analyzing cybersecurity news and **improve the consistency and accuracy** of threat classification. While the system shows potential, the current **low accuracy**, particularly during the evaluation phase, highlights the need for **further development**. The model can understand the **context and language** of cybersecurity news, but it still needs more **improvement** to ensure reliable and consistent threat detection. Moving forward, improving the model's **accuracy** remains a **crucial focus**, which will enhance its **practical application** in real-world cybersecurity scenarios.

Limitations and Challenges

- **Accuracy Issues:** Low accuracy due to:
  - Insufficient or imbalanced dataset.
  - Overfitting and model's difficulty generalizing to new attack types.

- **Complex Attack Descriptions:** Variability in attack styles and terminologies.

- **Data Quality & Size:** Inadequate coverage of attack techniques, poor learning of certain attack behaviors.

- **Language Limitation:** Currently only supports English, restricting global usability.

Future Work

- **Expand Dataset:** Include more types of attacks, descriptions, and tough examples for robustness.
- **Fine-tune Model:** Use more specific cybersecurity data for better understanding.
- **Multilingual Support:** Add support for different languages to broaden system’s applicability.
- **Integration with Live Threat Feeds:** Real-time classification of new threats by adapting to evolving attack methods.

Thank You