London, UK
September 02 - September 04, 2024

AI4COLLAB: An AI-based Threat Information Sharing Platform

C.Dalamagkas, D. Asimopoulos, P. Radoglou-Grammatikis, Nikolaos Maropoulos, Thomas Lagkas, Vasileios Argyriou, Gohar Sargsyan and Panagiotis Sarigiannidis

MetaMind Innovations P.C, Greece

AI4CYBER – Artificial Intelligence for next generation CYBERsecurity

Funded By the European Union

Authors & Contributors

Dimitrios-Christos Asimopoulos

Panagiotis Radoglou Grammatikis
Panagiotis Sarigiannidis, Nikolaos Maropoulos

Thomas Lagkas

Christos Dalamagkas

Vasileios Argyriou

Gohar Sargsya

Dimitrios-Christos Asimopoulos

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101070450 (AI4CYBER) .

Presentation Structure

Introduction

1

- Introduction
- Related Work
- Contributions

Main Part

2

- AI4COLLAB Architectural Design
- AI4COLLAB Detection Methods
- Evaluation Analysis

Conclusions

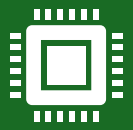
3

- Sum up important key points
- Future Work
- Q/A

Introduction, Related Work & Contributions

Introduction

Cyber threats are evolving rapidly, utilizing advanced technologies such as Artificial Intelligence (AI) to become more complex and sophisticated. This evolution includes not just targeting critical infrastructure but also expanding to areas like supply chains and IoT devices.



Attackers are increasingly using social engineering techniques like phishing and spear-phishing, aimed at exploiting human psychology to gain unauthorized access to sensitive information.

Cyber Threat Intelligence (CTI) involves the proactive collection and analysis of information about potential and real-time security threats and vulnerabilities. This includes gathering data from sources like security research, hacking forums, and network monitoring to identify indicators of compromise (IoCs) and predict future threats.

Related Work

T. Chantzios et al.

- Discusses the selection of appropriate cyber-threat intelligence sharing platforms, comparing different platforms based on their features and interoperability

G. Sakellariou et al.

- Proposes a reference model for cyber threat intelligence (CTI) systems to enhance data security, anonymity, and efficiency

2019 **2020** **2022** **2023**

A. Ramsdale et al.

- Provides a comparative analysis of cyber-threat intelligence sources, formats, and languages, focusing on their efficiency in various cybersecurity use cases

S. Nikoletos et al.

- Focuses on automated sensitive data identification and anonymization techniques in cybersecurity, addressing privacy concerns in data sharing

Contributions

Automated Detection of Sensitive Information	We utilize and compare various AI techniques to automatically detect sensitive information.
Anonymization of Sensitive Information	We automatically anonymize the sensitive information, while ensuring that the CTI data still conforms to the MISP format.
Integration of multiple CTI sharing platforms	Our proposed architecture showcases how two popular CTI platforms, Malware Information Sharing Platform (MISP)and OpenCTI, can be integrated and utilized to widen the range of CTI collection.

AI4COLLAB Architectural Design

AI4COLLAB Architecture Design

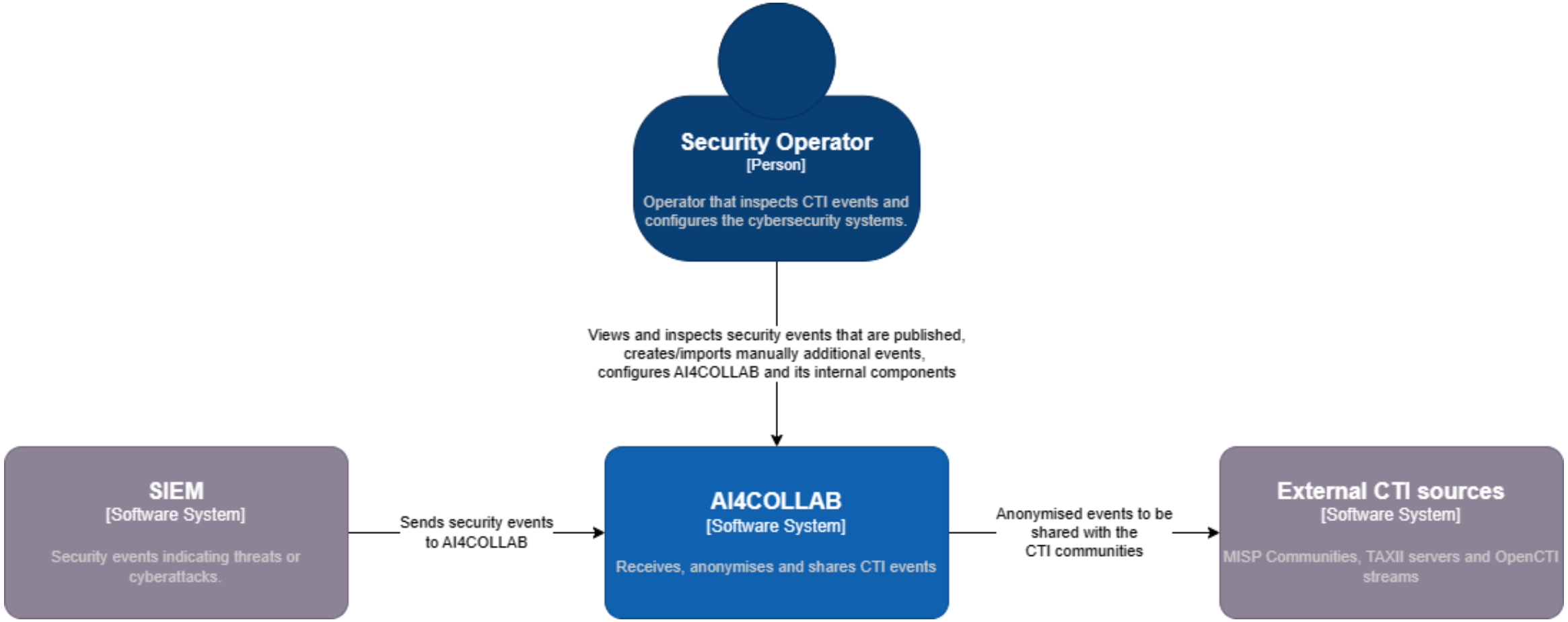
The AI4COLLAB architecture has been defined based on the **C4 model methodology**. The C4 model provides hierarchical abstractions to describe and visualize software architecture.

The C4 Methodology defines 4 types of diagrams

- **System Context Diagram**: Shows users and external entities interacting with the system.
- **Container Diagram**: This represents the system as a set of independent services that interact with each other.
- **Component Diagram**: Breaks down each container to detail components as function blocks performing specific tasks.
- **Code Diagram**: Describes the implementation of each component, utilizing UML diagrams and entity relationship diagrams.

A code diagram is not needed because of the simplicity of our custom code

System Context Diagram (1)



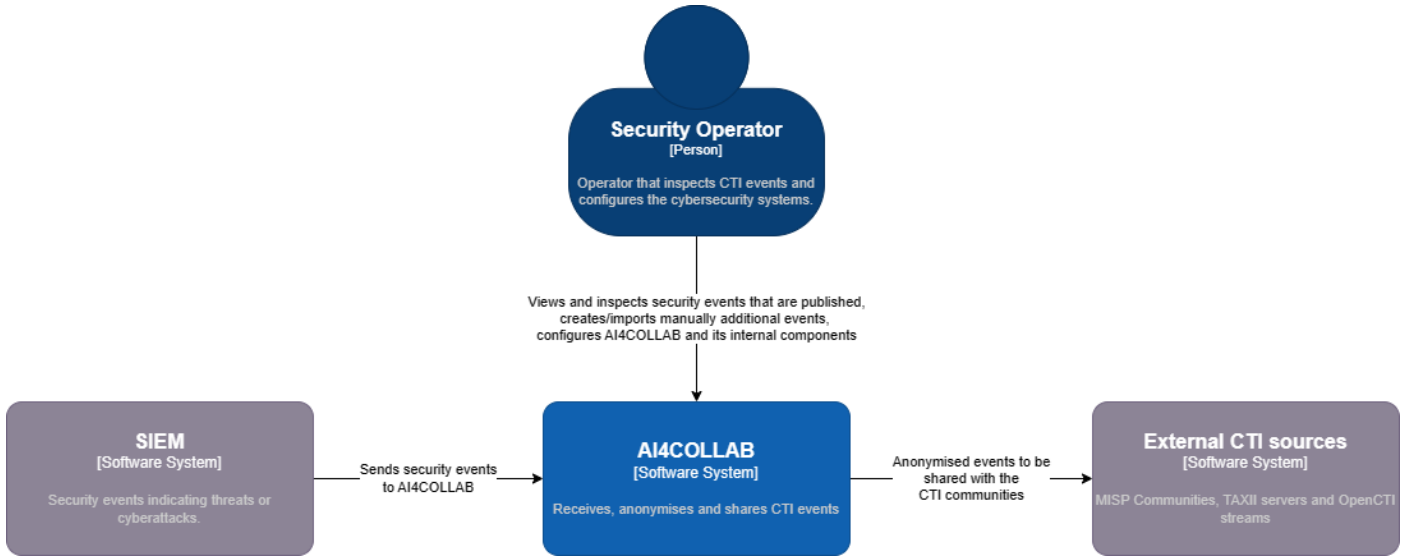
System Context Diagram (2)

Represents how AI4COLLAB interacts with users (Security Operators) and external entities (security systems, CTI communities).

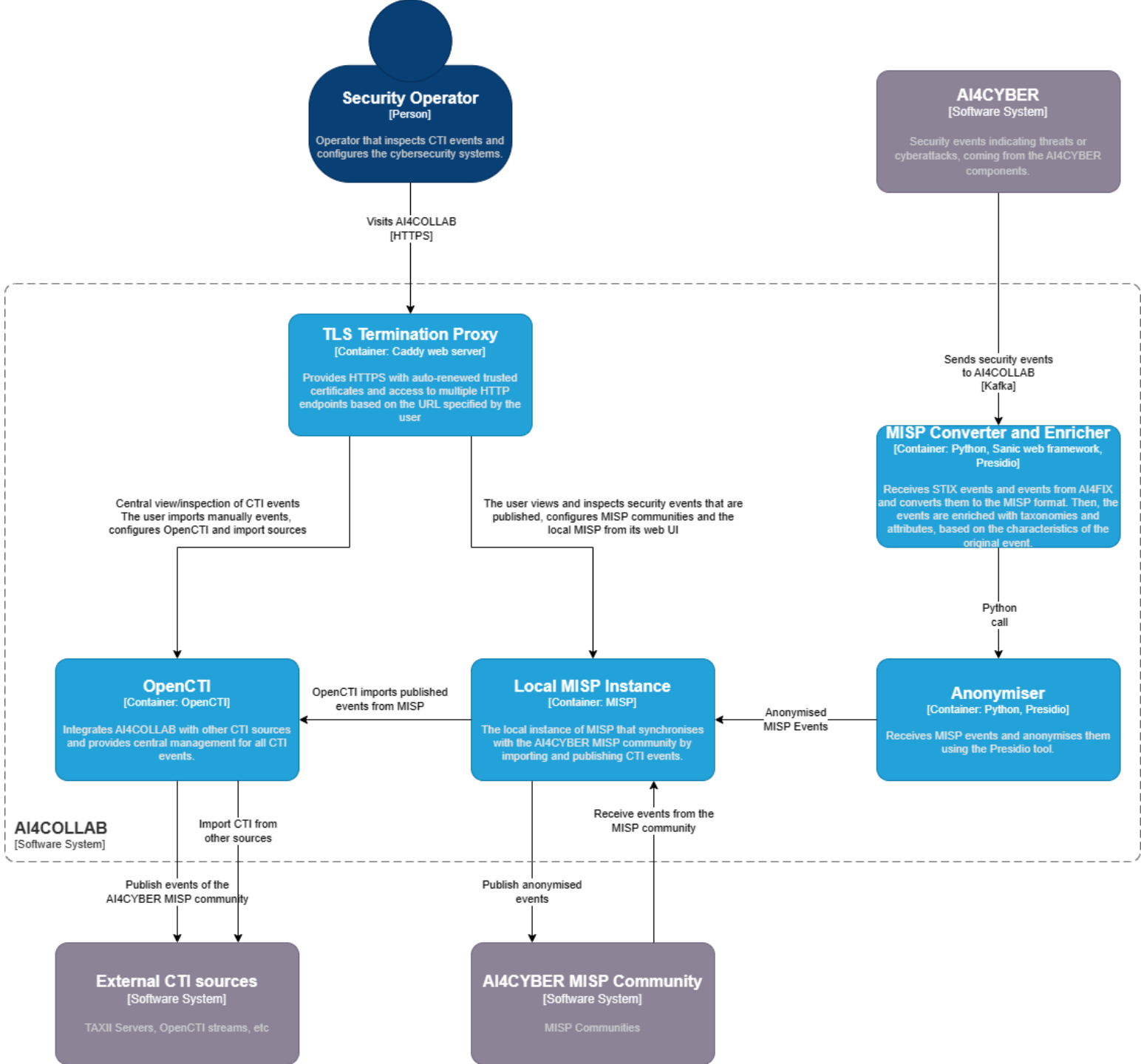
- **Users** are responsible for configuring the system and inspecting CTI data.
- **External Entities**
  1. SOAR Systems send security events to AI4COLLAB which describe identified threats and cyberattacks.
  2. CTI Sources include TAXII servers, CTI streams, and MISP instances that receive anonymized CTI data and reports from Ai4COLLAB.

Data Flow and Interactions

1. **Incoming Data**: AI4COLLAB receives security events from external security systems.
2. **Processing and Anonymization**: The system processes and anonymizes the incoming data to comply with privacy regulations.
3. **Data Sharing**: Share anonymized CTI data with external CTI sources to facilitate community-based threat intelligence.



Container Diagram (1)



Container Diagram (2)

The Container diagram analyzes the AI4COLLAB system into micro services that operate independently as separate system services. AI4COLLAB consists 5 containers.

MISP Converter and Enricher

Serves as the entry point for receiving security events for anonymization

1. Converts incoming events into the MISP format
2. Enriches MISP events with relevant Indicators of Compromise and taxonomies to maximize information representation

AI-based Anonymizer

Process and anonymizes MISP events

1. Detects sensitive information using AI algorithms
2. Applies data masking to anonymize the identified sensitive information
3. Submits the anonymized security event to the local MISP instance

Local MISP instance

- A local installation of the MISP platform for community participation and threat inventory management
- Security operators can inspect, edit, and approve events for dissemination through a web-based GUI

OpenCTI

A full installation of the OpenCTI platform for enhanced CTI management

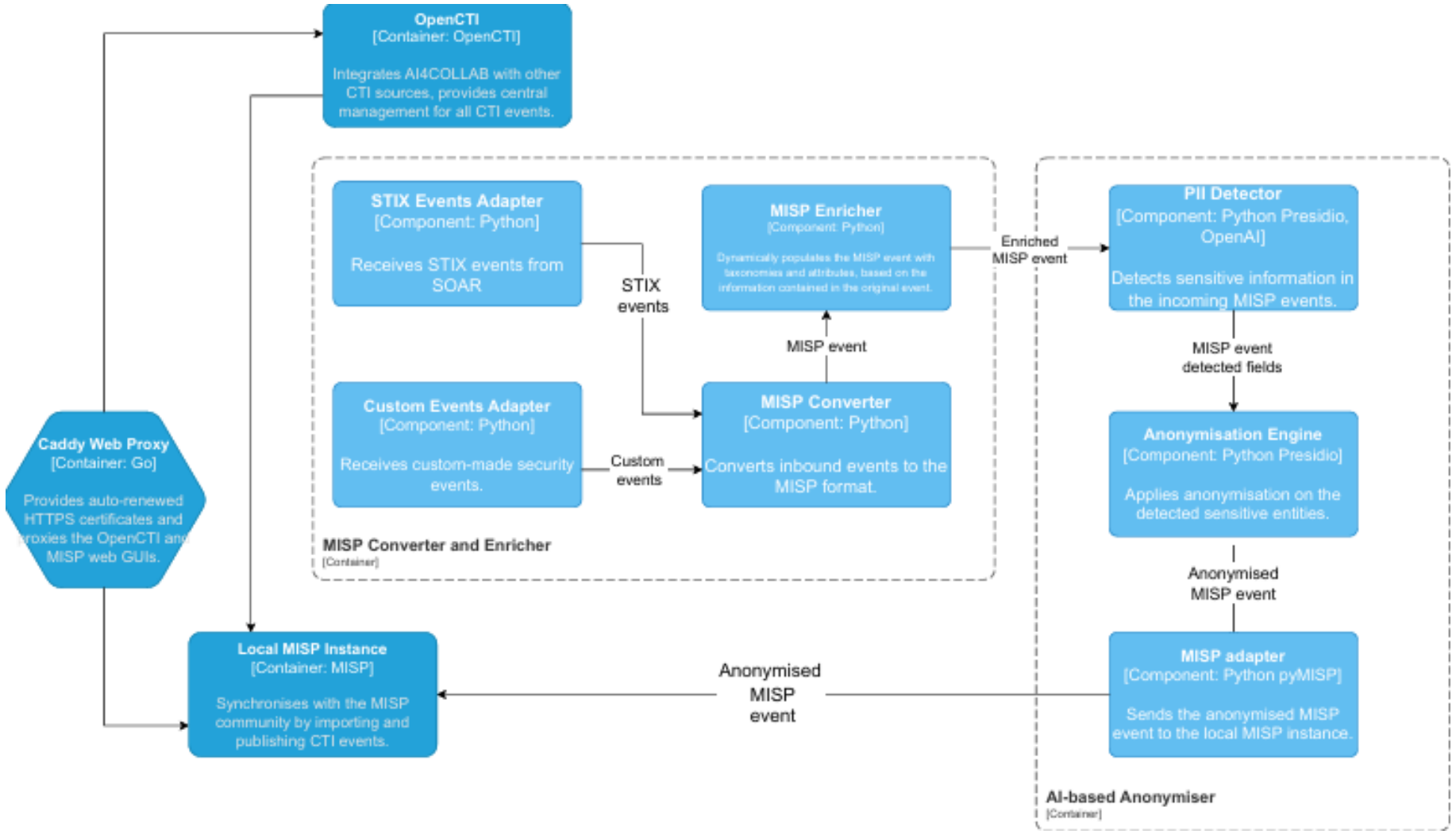
1. Imports MISP events periodically using a dedicated plugin to generate STIX reports
2. Provides a web-based GUI for CTI inspection and allows dissemination of intelligence through TAXII servers and CTI streams

TLS Termination Proxy

Based on the Caddy server, it serves as a reverse web proxy and TLS termination proxy

1. Provides HTTPS access to the MISP and OpenCTI web interfaces
2. Generates and renews HTTPS certificates through the Let's Encrypt service for secure internet access

Component Model (1)



Component Model (2)

The Component Diagram provides a detailed view of the implementation of two key services within AI4COLLAB: the MISP Converter & Enricher and the Ai-based Anonymizer.

Components of MISP Converter & Enricher

Event Adapters: Receives and process incoming events asynchronously

- Two parallel adapters handle STIX events and custom-made events
- Uses Apache Kafka to subscribe to distinct topics for new events

MISP Converter: Converts incoming events into MISP format

- Utilizes the MISP-STIX library for STIX events
- Custom code matches fields for custom events
- Validates and constructs a pyMISP object for each event

MISP Enricher: Enhances MISP events with additional taxonomies and attributes

- Uses misp-moduls service to retrieve information for Critical information in the events

Components of AI-based Anonymization

PII Detector: Detects sensitive information in enriched MISP events

- Can use multiple detection methods, such as ChatGPT(via OpenAI) and Presidio analyzer
- Outputs the original event and identified PII entities

Anonymization Engine: Applies data masking to sensitive information

- Replace sensitive data with wildcard characters(e.g., '#' or '*')
- Outputs anonymized MISP events

MISP adapter: Interacts with the local MISP instance to submit anonymized events

- Uses pyMISP library to save and publish events to MISP communities
- Can be configured to allow manual review of event before publication.

AI4COLLAB Detection Methods

AI4COLLAB Detection Methods

AI4COLLAB employs three methods for detecting Personally Identifiable Information (PII)

Microsoft Presidio

- A specialized tool for detecting and anonymizing sensitive information.

GPT-2 Model

- Utilized for text processing and detection.

ChatGPT

- Advanced AI model for identifying sensitive data.

Microsoft Presidio

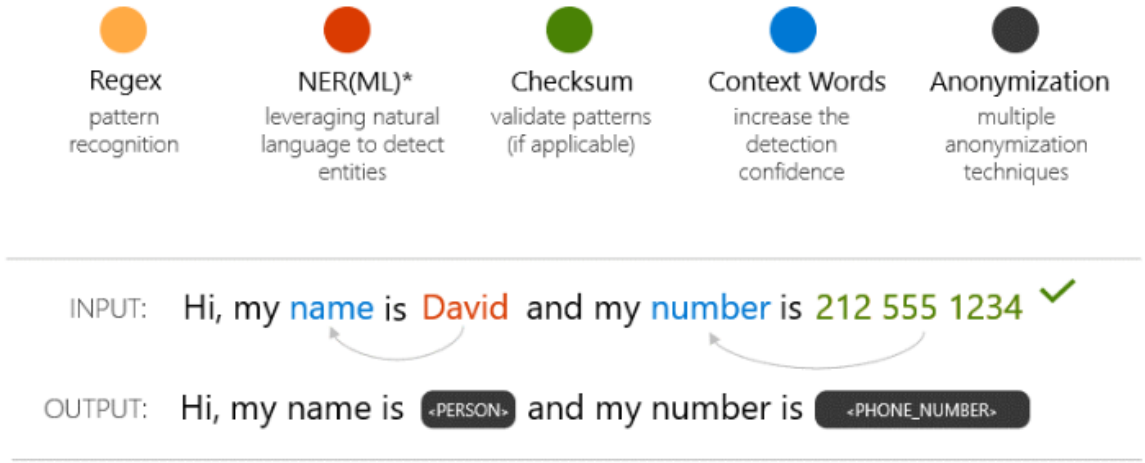
Presidio is designed to detect and anonymize a wide range of personal data types using machine-learning techniques.

Methodology

Predefined and Customizable Detectors: Use pattern recognition, checksum validation, and contextual analysis to identify sensitive data.

•**Anonymization Strategies**: Includes substitution, redaction, and generalization to obscure detected information.

Presidio Detection Flow



CREDIT_CARD	A credit card number is between 12 to 19 digits. https://en.wikipedia.org/wiki/Payment_card_number
CRYPTO	A Crypto wallet number. Currently only Bitcoin address is supported
DATE_TIME	Absolute or relative dates or periods or times smaller than a day.
EMAIL_ADDRESS	An email address identifies an email box to which email messages are delivered
IBAN_CODE	The International Bank Account Number (IBAN) is an internationally agreed system of identifying bank accounts across national borders to facilitate the communication and processing of cross border transactions with a reduced risk of transcription errors.
IP_ADDRESS	An Internet Protocol (IP) address (either IPv4 or IPv6).
NRP	A person's Nationality, religious or political group.
LOCATION	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains
PERSON	A full person name, which can include first names, middle names or initials, and last names.
PHONE_NUMBER	A telephone number
MEDICAL_LICENSE	Common medical license numbers.
URL	A URL (Uniform Resource Locator), unique identifier used to locate a resource on the Internet

Large Language Models (LLMs) for Data Privacy

- LLMs, such as those in the GPT family, offer a novel approach to data privacy and anonymization.
- These models are trained on diverse datasets, enabling them to understand various environments, idiomatic expressions, and semantic details.

Advantages of LLMs for Data Privacy

- **High Accuracy:** LLMs can detect subtle nuances in language, making them highly accurate in identifying sensitive information.
- **Flexibility:** Capable of adapting to various contexts and types of data, making them significant tools for different anonymization needs.
- **Maintaining Text Quality:** By providing contextually appropriate replacements, LLMs ensure the anonymized text remains of high quality, readable, and coherent.

Anonymization Process with LLMs

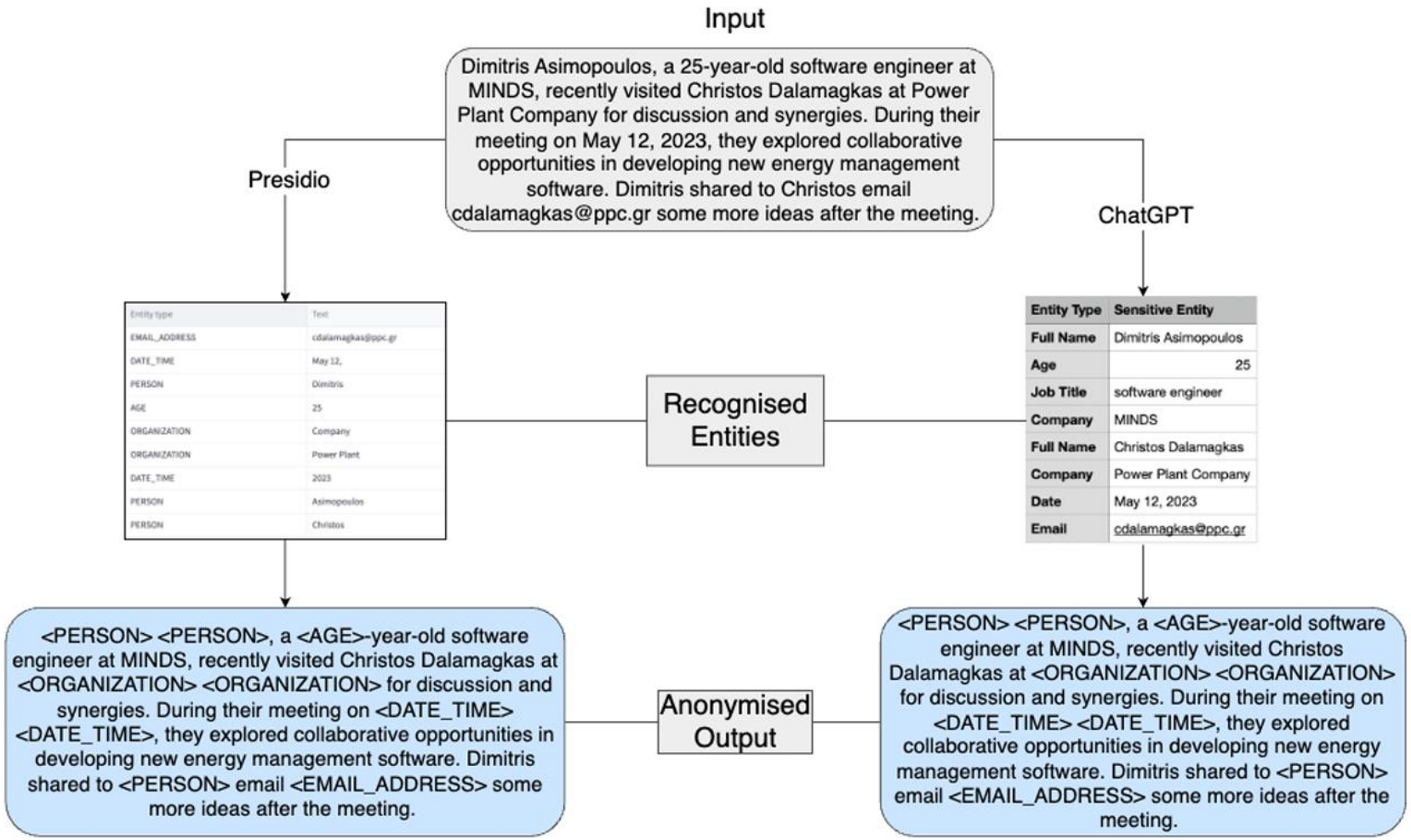
Fine-Tuning with Annotated Datasets:

- The process involves using annotated datasets containing sensitive information to fine-tune the GPT model.
- This enables the model to accurately identify specific categories of sensitive data.

Generating Contextual Replacements:

- GPT models generate replacements that are not only anonymized but also suitable within the context of the original text.
- Maintains the legibility and natural flow of the text, enhancing privacy without sacrificing clarity.

Anonymization Example



Evaluation Analysis

Evaluation- Data

Two experiments were carried out:

1. On a publicly available NER dataset (CONNL-2003) annotated with Part of Speech(POS) and Tag for each word. In this experiment Presidio and GPT-2 model were tested.
2. On MISP events were Presidio and ChatGpt were tested.

Tag	Description
I-LOC	Inside Location
B-ORG	Beginning of Organisation
O	Other
B-PER	Beginning of Person
I-PER	Inside of Person
I-MISC	Inside Miscellaneous
B-MISC	Beginning Miscellaneous
I-ORG	Inside of Organisation
B-LOC	Beginning of Location

```
Japan NNP B-NP B-LOC
began VBD B-VP O
the DT B-NP O
defence NN I-NP O
of IN B-PP O
their PRP$ B-NP O
Asian JJ I-NP B-MISC
Cup NNP I-NP I-MISC
title NN I-NP O
with IN B-PP O
a DT B-NP O
lucky JJ I-NP O
2-1 CD I-NP O
win VBP B-VP O
against IN B-PP O
Syria NNP B-NP B-LOC
in IN B-PP O
a DT B-NP O
Group NNP I-NP O
C NNP I-NP O
championship NN I-NP O
match NN I-NP O
on IN B-PP O
Friday NNP B-NP O
. . O O
```

Evaluation – Evaluation Metrics

Experiment results were evaluated with several metrics:

Evaluation Metrics	Description	Formula
Recall	The **Recall** metric, sensitivity or true positive rate, is a performance measure in machine learning and statistics that evaluates the accuracy of a classification model.	$\text{Recall} = \frac{TP}{TP + FN}$
Precision	**Precision** measures the proportion of correctly predicted positives out of all instances predicted as positive.	$\text{Precision} = \frac{TP}{TP + FP}$
F1 Score	**F1 score** is a metric that captures the balance between true positive rate (TPR) and precision.	$\text{F1} = \frac{2 \times TP}{2 \times TP + FP + FN}$

Evaluation Results

EVALUATION METRICS - CONLL-2003 DATASET

Model	F1-score	Precision	Recall
Presidio	0.85	0.88	0.83
GPT-2	0.71	0.70	0.79

EVALUATION METRICS FOR MISP EVENTS

Model	F1-score	Precision	Recall
Presidio	0.75	0.90	0.82
ChatGPT	0.89	0.80	0.84

Presidio: Effective in detecting most entities with high recall; however, it has a tendency for more false positives in some scenarios.

GPT-2: Balanced performance but needs improvement in precision to reduce incorrect identifications.

ChatGPT: Achieves a good balance between precision and recall, making it a reliable choice for sensitive entity recognition.

Conclusion & Future Work

Conclusion

AI4COLLAB Platform Overview

- A Cyber Threat Intelligence (CTI) solution designed to integrate multiple CTI protocols and standards.
- Utilizes AI to enhance threat intelligence sharing among various stakeholders and operators.

Architecture Design Overview

- The platform is built on the C4 model approach, detailing the internal implementation and system components.

Anonymization with AI

- AI techniques are employed to detect and anonymize sensitive information.
- Promotes voluntary collaboration and strengthens trust among participants by ensuring privacy.

Future Work

Enhancing Anonymization techniques

- Focus on improving anonymization by exploring privacy-preserving techniques such as K-anonymity and Differential privacy
- Ensure compliance with evolving privacy laws while maximizing the utility of shared data.

Exploring Advanced LLM-Based Methods

- Plan to investigate more sophisticated large language models (LLMs) for detection, including GPT-4 and Google Gemini

Expand the platforms capabilities

Minds

UNIVERSITY OF WESTERN MACEDONIA

AI4CYBER

AI4CYBER – Artificial Intelligence for next generation CYBERsecurity

Thank you for your attention!

IEEE CSR Cyber Security and Resilience