Projektthema / Thesis

KI-basierte Generierung und Validierung von synthetischen SIEM Log-Daten

Ziel:

Ziel ist die Entwicklung eines LLM-basierten Generierungstools für synthetische SIEM-Logdaten, die zum Testen eines Open-Source-SIEM-Systems eingesetzt werden sollen. Logdaten, die konkrete Angriffsszenarien beinhalten sind dabei von besonderem Interesse. Im Bereich der Cybersicherheit sind öffentliche Daten aus der realen Welt aufgrund von Datenschutzbedenken selten verfügbar. Herkömmliche Anonymisierungsmethoden können sensible Informationen oft nicht vollständig schützen.

Initial soll recherchiert werden, welche KI-Methoden für die Generierung von SIEM-spezifischen Log-Daten bisher untersucht wurden und aussichtsreich für die gestellte Aufgabe sind und welche öffentlich verfügbaren Log-Daten für ein Finetuning von KI-Modellen anwendbar sind. Daraus abgeleitet wir ein SIEM Log Data Generator (SIEM LDG) inklusive eines wiederholbaren Workflow entwickelt, der ein oder mehrere LLM Modelle verwendet, um Log-Daten zu erzeugen. Basieren auf öffentlich verfügbaren Log-Daten und intern bereitgestellten Log-Daten soll der SIEM LDG trainiert werden. Die so erzeugten Log-Daten sind zu validieren und es ist zu untersuchen, welchen Einfluss verschiedene LLM Modelle auf die Ergebnisse haben.

Aufgaben:

- Recherche zum Stand der Technik von KI-basierten Log-Daten Generatoren
- Recherche öffentlich verfügbarer Log-Daten aus den Bereichen Cyber-Sicherheit, Security Operations Center (SOC), Security Information and Event Management (SIEM)
- Recherche potentiell geeigneter LLM Modelle
- Entwicklung einer Architektur und eines Workflow zur Generierung von Log-Daten mit Hilfe von LLM
- Finetuning der LLM-Modelle mit gegeben Log-Daten
- Validierung der generierten Log-Daten und Überprüfung der Güte des Outputs
- Optimierung des SIEM LDG durch den Vergleich verschiedener LLM Modelle und Optimierung des Prompting
- Dokumentation der Evaluationsergebnisse, der Architektur und Implementierung, Grenzen des Ansatzes, mögliche Erweiterungen

Ideen für Basis-Technologien (Environment):

- Linux, Docker/Kubernetes, Python/Go, REST-APIs, JSON/YAML, Git, MITRE ATT&CK

Ideen/Publikationen:

[1] H. S. Galadima, C. Doherty and R. Brennan, "Towards LLM-based Synthetic Dataset Generation of Cyber Incident Response Process Logs," 2024 Cyber Research Conference - Ireland (Cyber-RCI), Carlow, Ireland, 2024, pp. 1-4, doi: 10.1109/Cyber-RCI60769.2024.10939563.

[2] Marcel Bizon: „Synthetisierung von semantisch korrekten Log-Daten für das Training von KI-Modellen“, Bachelor Thesis, TH Köln 2026.

[3] Karsten Mandt: „Automatisierte Generierung von Integrationstests in Kubernetes-Umgebungen mithilfe von LLMs“, Bachelor Thesis, TH Köln 2026.

[4] A Review of the Use of Large Language Models to Generate Synthetic Cybersecurity Datasets: https://www.computer.org/csdl/proceedings-article/itnac/2025/11302633/2cGhWinIoo0

[5] SAGA: Synthetic Audit Log Generation for APT Campaigns: https://arxiv.org/abs/2411.13138

[6] SIEVE: : Generating a cybersecurity log dataset collection for SIEM event classification: https://dl.acm.org/doi/10.1016/j.comnet.2025.111330

[7] CyberLLM-FINDS 2025: Instruction-Tuned Fine-tuning of Domain-Specific LLMs with Retrieval-Augmented Generation and Graph Integration for MITRE Evaluation: https://arxiv.org/html/2601.06779v1

Level: Master

Sprache: Deutsch oder Englisch

Beginn: sofort

Studierendenprojekt in Kooperation von DN.lab (TH Köln) und Procyde
Interessenten wenden sich an Thomas Karl (karriere@procyde.com) oder Prof. Grebe (andreas.grebe@th-koeln.de) und senden ihren PSSO-Auszug mit.

AI-Based Generation and Validation of Synthetic SIEM Log Data

Objective:

The objective is the development of an LLM-based generation tool for synthetic SIEM log data to be used for testing an open-source SIEM system. Log data containing realistic cyberattack scenarios are of particular interest. In the field of cybersecurity, real-world data are rarely publicly available due to privacy concerns, and traditional anonymization methods often fail to sufficiently protect sensitive information.

Initially, a review will be conducted to identify which AI techniques have been explored for generating SIEM-specific log data, which approaches appear promising for this task, and which publicly available log datasets can be applied for fine-tuning AI models. Based on these findings, a SIEM Log Data Generator (SIEM LDG) will be developed, including a reproducible workflow that employs one or more LLM models to generate synthetic log data.

Using both publicly available and internally provided data, the SIEM LDG will be trained to produce realistic and diverse log samples. The generated data will then be validated, and the influence of different LLM models on data quality and realism will be analyzed.

Tasks:

- Research the current state of AI-based log data generation methods
- Identify publicly available cybersecurity, SOC, and SIEM log datasets
- Evaluate potentially suitable LLM models for the task
- Develop an architecture and workflow for log data generation using LLMs
- Fine-tune LLM models with available log data

- Validate generated log data and assess the quality and realism of outputs
- Optimize the SIEM LDG by comparing different LLMs and improving prompting strategies
- Document evaluation results, architecture, implementation details, limitations, and possible extensions

Suggested Base Technologies (Environment):

Linux, Docker/Kubernetes, Python/Go, REST APIs, JSON/YAML, Git, MITRE ATT&CK

References and Inspiration:

[1] H. S. Galadima, C. Doherty and R. Brennan, "Towards LLM-based Synthetic Dataset Generation of Cyber Incident Response Process Logs," 2024 Cyber Research Conference - Ireland (Cyber-RCI), Carlow, Ireland, 2024, pp. 1-4, doi: 10.1109/Cyber-RCI60769.2024.10939563.

[2] Marcel Bizon: „Synthetisierung von semantisch korrekten Log-Daten für das Training von KI-Modellen“, Bachelor Thesis, TH Köln 2026.

[3] Karsten Mandt: „Automatisierte Generierung von Integrationstests in Kubernetes-Umgebungen mithilfe von LLMs“, Bachelor Thesis, TH Köln 2026.

[4] A Review of the Use of Large Language Models to Generate Synthetic Cybersecurity Datasets: https://www.computer.org/csdl/proceedings-article/itnac/2025/11302633/2cGhWinIoo0

[5] SAGA: Synthetic Audit Log Generation for APT Campaigns: https://arxiv.org/abs/2411.13138

[6] SIEVE: : Generating a cybersecurity log dataset collection for SIEM event classification: https://dl.acm.org/doi/10.1016/j.comnet.2025.111330

[7] CyberLLM-FINDS 2025: Instruction-Tuned Fine-tuning of Domain-Specific LLMs with Retrieval-Augmented Generation and Graph Integration for MITRE Evaluation: https://arxiv.org/html/2601.06779v1

Level: Master

Language: German or English

Start: Immediately

Type: Student project in cooperation with DN.lab (TH Köln) and Procyde

Contact:

Interested students should contact Thomas Karl (karriere@procyde.com) or Prof. Grebe (andreas.grebe@th-koeln.de) and include their PSSO transcript