Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Cyberbezpieczeństwo (CBD)**
Specjalność: **CBD**

PRACA DYPLOMOWA
MAGISTERSKA

Zastosowanie dużych modeli językowych do wykrywania i naprawiania błędów bezpieczeństwa i podatności w kodzie aplikacji webowych

Patryk Fidler

Opiekun pracy
Dr. hab. inż. Maciej Piasecki

Słowa kluczowe: modele językowe, Sztuczna Inteligencja, statyczna analiza kodu

WROCŁAW 2024

STRESZCZENIE

The engineering thesis titled "Application of Large Language Models for Detecting and Fixing Security Bugs and Vulnerabilities in Web Application Code" focuses on the utilization of advanced language models, such as GPT-3.5, GPT-4, and if possible, Falcon, for automated detection and rectification of security bugs in the code of software and web applications.

The motivation for this work stems from the growing role of Large Language Models (LLMs) in various fields, including cybersecurity. In the context of these efforts, the possibility of using these models to detect and fix vulnerabilities such as XSS, SQL Injection, CSRF, Buffer Overflow, and the like is being investigated.

The starting point for this dissertation is the 2021 article "Can OpenAI Codex and Other Large Language Models Help Us Fix Security Bugs?" by Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, Brendan Dolan-Gavitt (https://arxiv.org/pdf/2112.02125v1.pdf). The authors emphasize the significant potential of these models, and this work aims to continue this research and broaden its scope.

Planned activities include preparing a dataset containing vulnerable source code databases, testing the error detection capabilities of OpenAI language models, and comparing these results with existing solutions offered by Snyk.

Particular emphasis will be placed on the use of soft-prompting and in-context learning techniques, which can help improve detection and results, even with limited resources. The work also envisages the implementation of an autonomous AI agent capable of analyzing code, performing security tests, and making decisions based on the results of these tests and context.

Optionally, the possibilities of error detection by open language models will be explored, and in the longer term, the possibilities of specializing models in the field of cybersecurity through fine-tuning. All these actions aim not only to understand but also to improve the capabilities of LLMs in the context of cybersecurity.

The main goal of the work is to increase awareness of the potential of large language

models in cybersecurity and to propose practical solutions that can help developers create more secure applications.

ABSTRACT

Praca inżynierska zatytułowana "Zastosowanie dużych modeli językowych do wykrywania i naprawiania błędów bezpieczeństwa i podatności w kodzie aplikacji webowych" koncentruje się na zastosowaniu zaawansowanych modeli językowych, takich jak GPT-3.5, GPT-4, a w razie możliwości Falcon, do automatycznego wykrywania i naprawiania błędów bezpieczeństwa w kodzie oprogramowania i aplikacji webowych.

Motywacja tej pracy wynika z rosnącej roli dużych modeli językowych (LLM) w różnych dziedzinach, w tym w cyberbezpieczeństwie. W kontekście tych działań, badana jest możliwość wykorzystania tych modeli do wykrywania i naprawiania podatności takich jak XSS, SQL Injection, CSRF, Buffer Overflow i tym podobne.

Punktem wyjścia dla pracy dyplomowej jest artykuł napisany w 2021 roku "Can OpenAI Codex and Other Large Language Models Help Us Fix Security Bugs?" - Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, Brendan Dolan-Gavitt (https://arxiv.org/pdf/2112.02125v1.pdf). Autorzy podkreślają znaczący potencjał tych modeli, a niniejsza praca ma na celu kontynuację tych badań i poszerzenie ich zakresu.

Planowane działania obejmują przygotowanie zbioru danych zawierającego podatne bazy kodu źródłowego, testowanie zdolności detekcji błędów przez modele językowe OpenAI, oraz porównanie tych wyników z istniejącymi rozwiązaniami, oferowanymi przez firmę Snyk.

Szczególny nacisk zostanie położony na wykorzystanie technik soft-prompting i in-context learning, które mogą pomóc w udoskonaleniu detekcji i wyników, nawet przy ograniczonych zasobach. Praca przewiduje również implementację autonomicznego agenta AI zdolnego do analizy kodu, wykonania testów bezpieczeństwa i podejmowania decyzji na podstawie wyników tych testów i kontekstu.

Opcjonalnie, badane będą możliwości detekcji błędów przez otwarte modele językowe, a w dalszej perspektywie, możliwości specjalizacji modeli w zakresie cyberbezpieczeństwa za pomocą fine-tuning'u. Wszystkie te działania mają na celu nie tylko zrozumienie, ale także poprawienie możliwości LLM w kontekście cyberbezpieczeństwa.

Głównym celem pracy jest zwiększenie świadomości na temat potencjału dużych modeli językowych w cyberbezpieczeństwie oraz proponowanie praktycznych rozwiązań, które mogą pomóc programistom w tworzeniu bardziej bezpiecznych aplikacji.

SPIS TREŚCI

WPROWADZNIE

Niniejszy szablon jest adaptacją szablonu opracowanego przez Pana Wojciecha Myszkę (patrz https://kmim.wm.pwr.edu.pl/myszka/projekty).

Został przetestowany w następujących narzędziach:

— Overleaf.com – wersja on-line; nie jest wymagana instalacja
— TeXStudio/TeXLive oraz TeXStudio/MiKTeX oraz TeXWorks/MiKTex

Użycie pełnej wersji może wymagać (np. dla zestawu TeXWorks/MikTeX):

— instalacji Python 2.7+ wraz z pakietem Pygments; ścieżki do obu narzędzi powinny być ustawione w zmiennej środowiskowej PATH – pakiet jest używany do kolorowania słów kluczowych w listingach języków programowania
— w środowisku Latex należy włączyć opcję -shell-escape – jest wymagana dla pakietu "minted" Latexa

CEL PRACY

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

ZAKRES PRACY

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

1. TYTUŁ ROZDZIAŁU I

W książce [**?**] . . . można znaleźć informacje o wielu rzeczach, np. o . . .

1.1. PODROZDZIAŁ

Mój pies mnie rucha

2. TYTUŁ ROZDZIAŁU II

kEZU CZY TO NIGDY NIE BEDZIE DOBRZE DZIALAC

2.1. RYSUNKI

Na rysunku **??** ...

2.1.1. Dwa rysunki obok siebie

Na rysunkach **??** i **??** ...

2.2. TABELE

W tabeli **??** ...

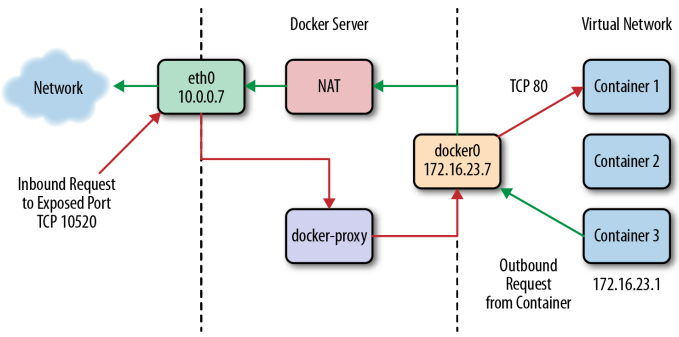
2.2.1. Równania

$$\sum_{i=1}^{\infty} a_i \tag{2.1}$$

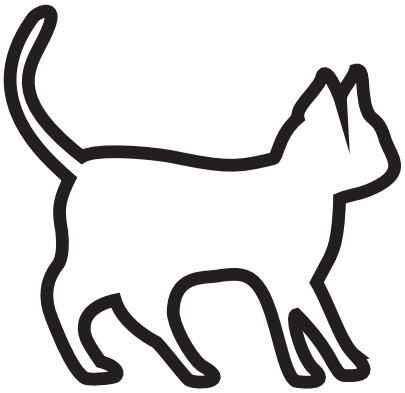
W równaniu **??** ...

2.3. LISTINGI

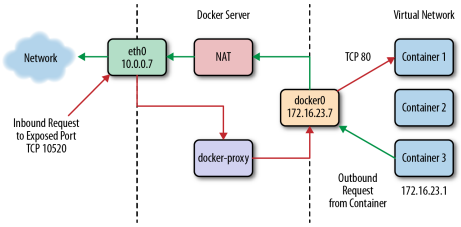
Na listingu **??** ...



Rys. 2.1. Sieć dokera



Rys. 2.2. Lewy rysunek



Rys. 2.3. Prawy rysunek

Tabela 2.1. Tytuł tabeli (patrz dodatek **??**)

Pierwszy	Drugi	Trzeci
Pierwszy	Drugi	Trzeci

Listing 2.1: Język C

PODSUMOWANIE

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam congue neque id dolor.

BIBLIOGRAFIA

[1] Docker Inc., *Compose file version 3 reference*, `https://docs.docker.com/compose/compose-file/`. Ost. dost. 12 listopada 2018.

SPIS RYSUNKÓW

SPIS LISTINGÓW

SPIS TABEL

Dodatki

A. DODATEK 1

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.