netidee

PROJEKTE

WebSecBot

Zwischenbericht | Call 18 | Projekt ID 6878

Inhalt

1 Einleitung

Das Projekt **WebSecBot** verfolgt das Ziel, ein Browser-Addon zu entwickeln, das Webentwickler:innen bei der Identifizierung und Behebung von Sicherheitslücken in Webanwendungen unterstützt. Durch den Einsatz von großen Sprachmodellen, Finetuning und einem Retrieval-Augmented Generation (RAG)-System bietet der WebSecBot eine semi-automatisierte Analyse von Webseiten und leitet Nutzer gezielt durch sicherheitsrelevante Prüfungen.

Im bisherigen Projektverlauf wurden wesentliche Meilensteine erreicht: Die erste Version des Browser-Addons wurde fertiggestellt und erfolgreich getestet, verschiedene lokale Sprachmodelle wurden evaluiert, und ein RAG-System wurde auf Basis der OWASP-Dokumentation implementiert. Diese Fortschritte bilden die Grundlage für die weiteren Optimierungen des Systems, um Benutzer:innen mit wenig Sicherheitserfahrung effizient durch komplexe Sicherheitstests zu führen.

Jedoch kam es zu Verzögerungen, da eine Mitarbeiterin nicht mehr für das Projekt zur Verfügung stand. Dies führte zu einer Umverteilung von Arbeitsstunden zwischen den verbleibenden und neuen Teammitgliedern und zu einem Rückstand im Zeitplan. Zudem hat sich gezeigt, dass das Finetuning der verwendeten Modelle mehr Zeit in Anspruch nimmt als ursprünglich geplant. Aus diesem Grund haben wir den Projektplan angepasst (siehe Kapitel 4), den Zwischenbericht später als ursprünglich geplant erstellt und beantragen eine kostenneutrale Verlängerung der Projektdauer von 12 auf 18 Monate, um alle geplanten Arbeiten erfolgreich abschließen zu können.

Die nächsten Arbeitspakete konzentrieren sich auf die Evaluierung des Systems in realistischen Szenarien wie dem OWASP Juice Shop, die Verbesserung der Benutzeroberfläche und die Zusammenarbeit mit externen Tester:innen zur Qualitätssicherung.

2 Status der Arbeitspakete

2.1 Arbeitspaket 1 – *Detailplanung und Formales am Projektstart*

Arbeitspaket 1 wurde zu Beginn des Projekts mit der Prüfung und Unterzeichnung des Fördervertrags abgeschlossen. Die Projektplanung wurde auf Basis der Vorlage erstellt und an den Fördergeber übermittelt. Im Rahmen dieses Arbeitspakets wurde außerdem die Projektwebsite mit den ersten Inhalten gefüllt und ein initialer Blogeintrag veröffentlicht, der das Projektteam sowie die Ziele des Projekts vorstellt.

2.2 Arbeitspaket 2 – *Evaluierung lokaler KI-Modelle & Finetuning*

Um die Unabhängigkeit von externen Dienstleistern zu ermöglichen, wurden verschiedene lokale Sprachmodelle evaluiert, um das Modell auszuwählen, das am besten für Web-Sicherheitstests geeignet ist. Im Test standen GPT-4(o) als Referenzmodell, Vicuna 13, Llama 3(.1) und Mixtral, die hinsichtlich ihres inhärenten Wissens in Bezug auf Sicherheitstests sowie ihrer allgemeinen Leistungsfähigkeit in verschiedenen Szenarien miteinander verglichen wurden.
Nach einer gründlichen Evaluierung erwies sich Llama 3.1 als das am besten geeignete Modell im Hinblick auf die Verarbeitung sicherheitsrelevanter Informationen und die Effizienz bei der Beantwortung komplexer Fragen. Auf Grundlage dieser Ergebnisse wurde entschieden, Llama 3.1 in den nächsten Schritten als Kernmodell zu verwenden.
Zusätzlich wurden im Zuge dieses Arbeitspakets die relevanten Quellen für das spätere RAG-System recherchiert. Der Fokus lag auf der OWASP-Dokumentation, einschließlich der OWASP Top Ten, des Web Security Testing Guide und der Cheatsheet Series, um sicherzustellen, dass das System über eine fundierte Wissensbasis verfügt.

Außerdem wurde das Llama 3.1 Modell in ein Retrieval-Augmented Generation (RAG)-System integriert. Dabei wurde das zuvor recherchierte sicherheitsrelevante Wissen, insbesondere OWASP-Ressourcen, in das RAG-System eingebunden. Dies ermöglichte es dem Modell, detailliertere und präzisere Vorschläge für sicherheitsrelevante Aufgaben zu machen.
Das RAG-System wurde daraufhin evaluiert, und die Ergebnisse zeigten, dass durch die Einbindung externer Wissensquellen signifikant an Leistung gewann. Diese Verbesserung bestätigte den Nutzen der Kombination aus lokalem Modell und externem Wissenszugang durch ein RAG-System.

2.3 Arbeitspaket 3 – *Server-Setup*

Für das Server-Setup wurde eine stabile und flexible Infrastruktur geschaffen, die die Ausführung von AnythingLLM und das darin integrierte Retrieval-Augmented Generation (RAG)-System unterstützt. Aktuell läuft die AnythingLLM-Instanz erfolgreich auf unserem Server, wobei für die Inferenz des Modells temporär Groq verwendet wird. Diese Lösung ermöglicht es uns, Testläufe unter realistischen Bedingungen durchzuführen und gleichzeitig die Systemleistung zu evaluieren. Langfristig planen wir jedoch, auf ein lokal gehostetes Modell umzusteigen, um vollständige Unabhängigkeit von externen Dienstleistern sicherzustellen und die Leistung weiter zu optimieren. Die Tests mit der aktuellen Konfiguration liefern wertvolle Erkenntnisse, die in die weiteren Entwicklungs- und Optimierungsschritte einfließen werden.

2.4 Arbeitspaket 4 – *Browser-Addon & LLM-Schnittstelle*

Im Rahmen dieses Arbeitspakets lag der Schwerpunkt auf der Entwicklung einer ersten Version des WebSecBots, welcher als Browser-Addon implementiert wurde. Das Addon wurde in React entwickelt und bietet eine Oberfläche, mit den Benutzer:innen Webseiten auf potenzielle Sicherheitslücken analysieren können. Zunächst wurde das Addon mit Unterstützung der OpenAI API umgesetzt, um schnell einen funktionierenden Prototypen zu erreichen.

Ein zentraler Bestandteil dieses Arbeitspakets war die Auswahl und Integration einer geeigneten Schnittstelle für die Sicherheitsanalysen. Hierbei wurden verschiedene Chatbot-Benutzeroberflächen evaluiert. In der aktuellen Version können Benutzer:innen innerhalb der WebSecBot-Oberfläche Sicherheits-Checklisten abarbeiten und erhalten konkrete Handlungsempfehlungen zur Verbesserung der Webanwendungssicherheit.

2.5 Arbeitspaket 5 – *Prompt-Engineering für interaktive Websecurity-Anwendungen*

Im Arbeitspaket 5 befinden wir uns aktuell in der Phase das WebSecBot-Modell auf interaktive Websicherheitsanwendungen zuzuschneiden. Dies umfasst die Recherche von vorhandenen Texten und Beispielen, die in den Kontext des verwendeten LLM eingebettet werden sollen.

Ein zentraler Aspekt dieses Arbeitspakets ist die Definition von System-Prompts, die unserem WebSecBot helfen sollen, Websicherheitsprüfungen basierend auf standardisierten Sicherheitsrichtlinien, wie den OWASP-Guidelines, durchzuführen. Wir analysieren verschiedene Möglichkeiten, wie das Modell gezielt auf sicherheitsrelevante Eingaben reagiert und dabei auf potentielle Gefahren hinweist. Hierbei ist es notwendig, den bestehenden Filter des LLM zu bewerten und Mechanismen zu finden, um sicherzustellen, dass sicherheitskritische Informationen korrekt verarbeitet und ausgegeben werden.

Derzeit testen wir unterschiedliche Ansätze für die System-Prompts und arbeiten daran, die bestmögliche Strategie zu implementieren. In den nächsten Schritten erfolgt die Integration der fertigen Prompts in das Browser-Addon, um eine reibungslose Nutzung in Echtzeitanwendungen zu gewährleisten.

2.6 Arbeitspaket 6 – *Testing und Evaluierung*

In diesem Schritt konzentrieren wir uns auf die funktionale Prüfung des Browser-Addons und die inhaltliche Bewertung des Systems. Dazu setzen wir das Addon in realen Szenarien ein, wie dem OWASP Juice Shop, um dessen Effektivität bei der Identifikation von Sicherheitslücken zu testen. Ziel ist es, sicherzustellen, dass der WebSecBot Nutzer:innen ohne tiefgehende Erfahrung mit Websecurity effektiv durch Sicherheitsprüfungen führt und hilfreiche Vorschläge zur Verbesserung der Sicherheit von Webanwendungen bietet. Die Ergebnisse dieser Tests fließen direkt in die weiteren Optimierungen des Systems ein.

2.7 Arbeitspaket 7 – *Dokumentation und Formales am Projektende*

Dieses Arbeitspaket startet aufgrund der erwähnten Verzögerung im Juni 2025.

3 Umsetzung Förderauflagen

Das Projekt hat keine speziellen Förderauflagen

4 Zusammenfassung Planaktualisierung

Alle Anpassungen des Plan-Excels kurz zusammengefasst

Arbeitspaket 2 – Evaluierung lokaler KI-Modelle & Finetuning:
Die Laufzeit dieses Arbeitspakets wird bis Jänner 2025 verlängert, um zusätzliche Abstimmungen zu ermöglichen. Diese sind notwendig, um die Funktionalität der Modelle in realen Szenarien zu evaluieren und die Ergebnisse optimal in das System zu integrieren.

Arbeitspaket 3 – Server-Setup:
Aufgrund unerwarteter Leistungsprobleme mit unserem aktuellen Server mussten wir nach einer geeigneten Alternative suchen. Derzeit wird die AnythingLLM-Instanz über eine Groq-Hardwarelösung betrieben, um die Systemleistung während der Test- und Evaluierungsphasen sicherzustellen. Langfristig ist geplant, die Verarbeitung wieder vollständig auf lokaler Infrastruktur auszuführen. Dieses Arbeitspaket wird bis März 2025 verlängert, sodass die finale Umstellung auf die lokale Infrastruktur nach Abschluss der Tests erfolgen kann.

Arbeitspaket 4 – Browser-Addon & LLM-Schnittstelle:
Dieses Arbeitspaket wird bis Ende November verlängert. Verzögerungen entstanden durch Probleme bei der Verbindung der AnythingLLM-Instanz mit dem Browser-Plugin. Die verbleibende Zeit wird genutzt, um die Integration erfolgreich abzuschließen und die Funktionalität zu testen.

Arbeitspaket 5 – Prompt-Engineering für interaktive Websecurity-Anwendungen:
Da wesentliche Vorarbeiten in den vorhergehenden Arbeitspaketen mehr Zeit als geplant in Anspruch genommen haben, wird dieses Arbeitspaket bis März 2025 verlängert. Dies ermöglicht es, die Prompts umfassend zu testen und für die Integration in das System zu optimieren.

Arbeitspaket 6 – Testing und Evaluierung:
Die Laufzeit dieses Arbeitspakets wird bis Mai 2025 verlängert. Die zusätzliche Zeit wird benötigt, um die Funktionalität des Systems unter realen Bedingungen mit professionellen Penetration-Testern gründlich zu testen und sicherzustellen, dass es den Anforderungen der Nutzer:innen entspricht und die Hilfestellungen tatsächlich die Sicherheit der Webanwendung erhöhen.

Arbeitspaket 7 – Dokumentation und Formales am Projektende:
Dieses Arbeitspaket startet aufgrund der erwähnten Verzögerungen im Juni 2025.

5 Öffentlichkeitsarbeit/ Vernetzung

Beschreibung der bereits erfolgten Öffentlichkeitsarbeit oder Vernetzung, bzw. Beschreibung des Plans künftiger Aktivitäten

Für das Projektende ist eine Präsentation des entwickelten Add-ons im Rahmen eines Security-Meetups bei SBA Research geplant. Diese Veranstaltung bietet eine ideale Gelegenheit, das Add-on einer breiteren Fachöffentlichkeit vorzustellen und den Austausch mit anderen Experten aus der IT-Sicherheitsbranche zu fördern.

6 Eigene Projektwebsite

Es wird keine eigene Webseite für das Projekt betrieben. Der WebSecBot wird als Browser-Addon entwickelt und steht derzeit während der Implementierungsphase noch nicht öffentlich zur Verfügung. Nach Abschluss des Projekts wird das Addon der Öffentlichkeit über die verfügbaren Browser-Addon-Stores zugänglich gemacht und weiterhin auf Servern der Uni Wien betrieben.