Exposé: LLM-unterstützte Threat-Modellierung mit DFDs

1. Motivation und Problemstellung

Die zunehmende Verbreitung von Künstlicher Intelligenz (KI) verändert die Arbeitswelt in vielen Bereichen. Besonders im der Industrie 4.0 ermöglichen digitale Systeme neue Möglichkeiten. Gleichzeitig bringen sie aber auch größere Herausforderungen in Bezug auf die IT-Sicherheit mit sich. Da Software immer komplexer wird, ist es wichtig, rechtzeitig Schwachstellen zu erkennen, um spätere Angriffe oder Systemausfälle zu verhindern. Ein bewährtes Mittel dafür ist das sogenannte Threat Modeling.

Was genau ist Threat Modeling?

Threat Modeling ist ein strukturierter Ansatz, bei dem Sicherheitsrisiken in IT-Systemen aufgedeckt werden. Ziel ist es, bereits in der frühen Entwicklungsphase potenzielle Bedrohungen sichtbar zu machen, damit diese nicht ausgenutzt werden können. Sicherheitsexpert:innen, Softwarearchitekt:innen oder Entwicklungsteams arbeiten häufig mit Threat Modeling. Ein essentielles Werkzeug in diesem Prozess sind Data Flow Diagrams (DFDs).

Was ist ein Data Flow Diagram (DFD)?

Ein DFD zeigt anschaulich, wie Daten in einem System fließen, wie bei Internet-of-Things (IoT). Es verdeutlicht, wer am Systems beteiligt ist, welche internen Abläufe stattfinden und wie Informationen zwischen den einzelnen Komponenten übertragen werden. Das Ergebnis ist ein übersichtliches Bild des Systems, das die Basis für gezielte Sicherheitsanalysen bildet.

2. Problemstellung und Leitfrage

Obwohl Threat Modeling ein äußerst wichtiges Instrument für die IT-Sicherheit ist, gestaltet sich der Prozess in der Praxis oft als ziemlich zeitaufwendig, fehleranfällig und stark von dem Wissen einzelner Personen abhängig. Neue KI-Technologien, vor allem Large Language Models (LLMs) wie ChatGPT, könnten hier wirklich helfen – zum Beispiel durch automatisch unterstützte Analyse von DFDs und die Entwicklung möglicher Bedrohungsszenarien.

Erste Studien (z. B. Mbaka & Tuma, 2024) zeigen vielversprechende Ansätze, weisen aber auch auf Herausforderungen hin: etwa Fehlalarme (False Positives), unklare Begründungen oder fehlender Kontext bei der Analyse komplexer Systeme (vgl. Yang et al., 2024).

Vor diesem Hintergrund untersucht die vorliegende Arbeit, inwiefern LLMs wie ChatGPT den Prozess des Threat Modeling mit Hilfe von DFDs unterstützen können. Dabei wird auch geprüft, wie eine Zusammenarbeit zwischen Mensch und Maschine gestaltet sein muss, um zuverlässig, verständlich und nützlich zu sein.

3. Ziel der Arbeit

Das Ziel dieser Arbeit ist es zu verstehen, wie verschiedene Fachleute – wie Experten für große Sprachmodelle, Softwareentwicklerinnen und Sicherheitsspezialisten – mit einem System zur Bedrohungsmodellierung arbeiten, das auf einem Sprachmodell basiert. Im Mittelpunkt steht die Frage, wie man sowohl den Prozess als auch die Ergebnisse der Bedrohungsanalyse verbessern kann und welche Bedingungen erfüllt sein müssen, damit eine vertrauenswürdige und nachvollziehbare Mensch-Maschine-Interaktion gelingt.

Dabei soll ein interaktives System entwickelt und getestet werden, das auf DFDs basiert und bei der Bedrohungsmodellierung unterstützt. Außerdem wird genauer angeschaut, wie die verschiedenen Fachgruppen diesen Prozess wahrnehmen, welche Erwatrungen sie haben und unter welchen Umständen sie ein solches System als hilfreich, vertrauenswürdig und verlässlich einstufen.

Zunächst wird in einem Gespräch mit unterschiedlichen Stakeholdern diskutiert, wie ein solches System effizient gestaltet werden kann. Auf Grundlage ihres Feedbacks, ihrer Anregungen und Ideen wird ein funktionaler Prototyp entwickelt. Anschließend wird analysiert, wie das System von den Beteiligten bewertet, erlebt und verstanden wird. Dabei stehen insbesondere die Nachvollziehbarkeit der LLM-Ausgaben und das Vertrauen an die Präzision im Vordergrund.

4. Methodik

Die Studie verfolgt das Ziel, den Einsatz eines LLMs wie ChatGPT 4.0 zur automatisierten Bedrohungsanalyse auf Basis von DFDs nach dem STRIDE-Modell zu erforschen. Dazu wird ein mehrstufiges methodisches Vorgehen gewählt, das explorativen Charakter hat und qualitative Erkenntnisse über die technische Umsetzbarkeit sowie die fachliche Plausibilität des Ansatzes liefern soll .

Zunächst wird ein vorbereitetes DFD, in Form einers Bildes, Chat-GBT gesendet. Daraufhin generiert das LLM STRIDE-basierte Bedrohungsvorschläge. Der Output wird von Expert:innen validiert und geben Feedback.

Zuerst wird ein vorbereiteter DFD als Bild an ChatGPT geschickt. Das LLM erstellt daraufhin Bedrohungsszenarien basierend auf dem STRIDE-Modell. Diese Ergebnisse werden anschließend von Fachleuten aus IT-Sicherheit, Softwareentwicklung und LLM-Technologie geprüft. Das Feedback fließt in die Entwicklung eines Web-Apps-Prototyps ein, mit dem Nutzer DFDs hochladen und automatische Analysen von ChatGPT erhalten können. Bei der Gestaltung der Eingabestrukturen werden verschiedene Prompt-Engineering-Strategien wie Zero-Shot, Few-Shot und Chain-of-Thought genutzt.

Der gesamte Prozess der Entwicklung und Bewertung findet in zwei Workshops statt. Im ersten Workshop wird das methodische Konzept einem interdisziplinären Publikum vorgestellt. Dabei gibt es eine klare Einführung in die wichtigsten Begriffe, das Ziel der Untersuchung, ähnliche wissenschaftliche Arbeiten sowie ein Beispielvideo, in dem ChatGPT 4.0 mit einem DFD gefüttert wird und eine Demo-STRIDE-Analyse zeigt. Dieses Video ist die Basis für eine offene Diskussion, bei der die Teilnehmenden kritisch hinterfragen, ob die Ergebnisse nachvollziehbar sind, ob wichtige Aspekte übersehen wurden und wie verständlich die Argumentation des Modells ist.

Im Verlaufe des Workshops arbeiten die Teilnehmenden in kleinen Gruppen, die jeweils nach ihrem Fachbereichen, LLM-Technologie, IT-Sicherheit und Softwareentwicklung unterteilt sind. Jede Gruppe bewertet die Bedrohungsanalyse, die sie erstellt hat, hinsichtlich der Qualität der Methode und der praktischen Umsetzbarkeit. Im Anschluss werden alle Einschätzungen in einer gemeinsamen Diskussion zusammengeführt. Dabei gibt es auch ein Live-Voting, bei dem unterschiedliche Aspekte bewertet werden. In diesem Rahmen sprechen wir über Fragen wie die Sinnhaftigkeit des Vorgehens, die spezifischen Herausforderungen in den jeweiligen Disziplinen und die Rolle des menschlichen Inputs. Weitere wichtige Themen sind das Vertrauen in KI-gestützte Systeme sowie die besten Wege, Informationen aufzubereiten, dass Nutzer mit dem Tool besser kommunizieren können.

Nach dem ersten Workshop, in dem wir das Konzept entwickelt haben, folgt nun ein zweiter, praktisch ausgerichteter Workshop. Dabei testen die Teilnehmenden einen Prototyp der Webanwendung. Der Schwerpunkt liegt dabei auf der Funktionalität, der Verständlichkeit des Konzepts und dem Nutzen des Outputs. Die Rückmeldungen werden qualitativ ausgewertet, indem wir beobachten, wie die Teilnehmenden das Tool verwenden, Interviews führen und spontane Meinungen sammeln. Dabei schauen wir besonders auf die Reaktionen, das Vertrauen in die Technologie, das Interaktionsverhalten und wie gut die Lösung insgesamt ankommt.

Das Ziel dieser Untersuchung ist nicht, verschiedene Tools oder Methoden direkt zu vergleichen. Stattdessen geht es vor allem darum herauszufinden, ob und wie der gesamte Ablauf funktionieren kann – vom Eingeben eines DFD über die automatische Analyse mit STRIDE bis hin zur menschlichen Bewertung und Nutzung in der Praxis. Es handelt sich hier um eine explorative Studie, die Hinweise darauf geben soll, wie sich KI-Modelle wie LLMs in sicherheitsrelevanten Softwareentwicklungsprozessen integrieren lassen könnten.

5. Relevante Studien

In den letzten Jahren haben mehrere Studien das Potenzial und die Grenzen von LLMs Mbaka und Tuma (2024)[1] zeigen, dass LLMs zusammen mit DFDs die Einschätzung realistischer Bedrohungen verbessern können. Teams, die LLMs nutzen, erzielen zwar mehr richtige Treffer, aber haben jedoch auch deutlich mehr False Positives. Außerdem wurde deutlich, dass vor allem weniger erfahrene Nutzer:innen dazu neigen, die ersten Modellvorschläge ohne Prüfungen zu übernehmen – was das Risiko von Fehlbewertungen erhöht.

Yang und Kolleg:innen (2024)[2] haben sich mit der automatischen Durchführung von Threat Modeling beschäftigt. Dabei wurden DFDs automatisch aus Quellcode und Dokumenten erstellt und für die Bedrohungsanalyse genutzt. Die Ergebnisse zeigen: LLMs können typische Angriffsmethoden zuverlässig erkennen und Bedrohungsszenarien generieren. Allerdings geht bei komplexen Systemen oft die Präzision verloren, vor allem durch fehlerhafte DFDs, mangelndes Fachwissen oder fehlende Erklärbarkeit.

[1] Vgl. W. Mbaka; K. Tuma: Usefulness of data flow diagrams and large language models for security threat validation: a registered report, auf: arXiv.org, 15. August 2024.
[2] Vgl.S. Yang; T. Wu; S. Liu; D. Nguyen; S. Jang; A. Abuadbba: THREATMODELING-LLM: Automating Threat Modeling using Large Language Models for Banking System, auf: arXiv.org, 26. November 2024.

Mollaeefar und Team (2024)[3] haben mit dem Tool PILLAR ein System bewertet, das auf LLMs basiert und automatisch Datenschutzbedrohungen erkennt. Das System analysiert DFDs, ordnet Risiken zu und nutzt bekannte Sicherheitsstandards wie GDPR oder ISO 27001. Kritisch sind jedoch das begrenzte Verständnis für den Kontext, die fehlende Nachvollziehbarkeit und manchmal unpassende Sicherheitstipps.

Eine frühere Studie von Scandariato und Kollegen (2015)[4] untersuchte, wie Informatik-Studierende die STRIDE-Methode manuell anwenden. Dabei wurde deutlich, dass STRIDE grundsätzlich lernbar und nutzbar ist, aber viel Zeit kostet. Die durchschnittliche Erkennungsrate lag bei unter 80 %, wobei bestimmte Bedrohungskategorien öfter übersehen wurden. Außerdem zeigte sich, dass mehr Zeitinvestition nicht automatisch zu besseren Ergebnissen führt.

Tuma und Mbaka (2022) haben sich außerdem mit dem Einfluss menschlicher Faktoren wie Geschlecht, Herkunft oder Nationalität auf die Qualität von Bedrohungsanalysen beschäftigt. Auch wenn die geplante Studie nie umgesetzt wurde, regt das Thema dazu an, darüber nachzudenken, wie Unterschiede in Wahrnehmung in sicherheitskritischen Prozessen berücksichtigt werden können.[5]

[3] Vgl. M. Mollaeefar; A. Bissoli; S. Ranise: PILLAR: An AI-POWERED PRIVACY THREAT MODELING TOOL, in: Department of Mathematics, University of Trento, Italy, 11. Oktober 2024.
[4]Vgl. R. Scandariato; K. Wuyts; W. Joosen: A descriptive study of Microsoft's threat modeling technique, am: 1. Juni 2015.
[5]Vgl. K. Tuma; W. Mbaka: Human Aspect of Threat Analysis: A Replication, auf: arXiv.org, 2. August 2022.