**LLM-unterstützte Threat-Modellierung mit DFDs**

**1. Motivation und Problemstellung**

Die zunehmende Verbreitung von Künstlicher Intelligenz (KI) führt zu zahlreichen grundlegende Veränderungen in der Arbeitswelt. Ob in der industriellen Produktion, im Gesundheitswesen, im Bildungsbereich oder in der öffentlichen Verwaltung, KI-Systeme übernehmen immer mehr Aufgaben, die früher ausschließlich von Menschen ausgeführt wurden. In der sogenannten Industrie 4.0 wird diese Entwicklung besonders sichtbar, in den vernetzten Systemen, die Maschinen, Sensoren, Datenplatformen und Software miteinander verknüpfen. Einerseits führt die stetig steigende Nutzung solcher Technologien zu beträchtlichem Effizienzgewinn und neuen Innovationspotenzialen.

Andererseits bringen sie aber auch größere Herausforderungen in Bezug auf die IT-Sicherheit mit sich. Vor allem in sicherheitskritischen Infrastrukturen und bei sensiblen Datenverarbeitungsprozessen stellen diese Entwicklungen große Herausforderungen dar. Die Vielschichtigkeit von solchen Systemen steigt, wodurch es immer anspruchsvoller wird, diese zu überblicken und abzusichern, was in Zeiten der fortschreitenden Digitalisierung und Vernetzung die IT-Sicherheit wiederum zu einem strategischen Erfolgsfaktor für Unternehmen macht, besondere bei komplexen, cloudbasierten oder IoT-basierten Anwendungen. Die klassische Herangehensweise, Software-Sicherheit nachträgliche zu testen oder erst nach auftreten von Vorfällen darauf zu reagieren, reicht nicht mehr aus. Deshalb wächst die Relevanz nach präventiven Maßnahmen wie das Testen von Sicherheitslücken bereits zu Beginn in der Planungs- und Entwicklungsphase eines IT-Systems. In einem solchen Verfahren ist das sogenannte Threat Modeling eine bewährte Mittel.

Was genau ist Threat Modeling?

Beim Threat Modeling handelt es sich um ein Prozess zur systematischen Identifikation und Bewertung von Sicherheitsrisiken in IT-Systemen. Dieser Prozess hilft bereits in der frühen Entwicklungsphase potenzielle Bedrohungen zu erkennen und entsprechende Gegenmaßnahmen zu entwicklen, damit diese nicht ausgenutzt werden können. In der Softwareentwicklungen arbeiten Sicherheitsexpert:innen, In der Softwareentwicklung verwenden Sicherheitsexpert:innen, Softwareexpert:innen oder Entwicklerteams häufig Threat Modeling, das oft als bedeutender Bestandteil im sicherheitsbewussten Programmierens angesehen wird. Unternehmen und Organisationen weltweit greifen auf verschiedene Methoden und Frameworks zurück, darunter STRIDE, DREAD oder PASTA, um sicherzustellen, dass ihre Systeme gegen bekannte und unbekannte Bedrohungen gesichert sind. Ein essentielles Werkzeug in diesem Prozess sind Data Flow Diagrams (DFDs).

Was ist ein Data Flow Diagram (DFD)?

Ein DFD zeigt anschaulich den Fluss von Daten innerhalb eines Systems, 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 der Systemsarchitekur, das die Basis für gezielte Sicherheitsanalysen bildet.

Obwohl Threat Modeling von großer Bedeutung für die IT-Sicherheit ist, erweist sich der Prozess in der Praxis oft als zeitaufwendig, fehleranfällig und stark abhängig vom domänenspezifischen Wissen einzelner Personen. Neue KI-Technologien, vor allem Large Language Models (LLMs) wie ChatGPT, könnten hier als aushelfen, beispielsweise durch automatisch unterstützte Analyse von DFDs oder durch die Entwicklung möglicher Bedrohungsszenarien und dem ausgeben von Vorschlägen für Gegenmaßnahmen.

Denn moderne IT-Architekturen bestehen aus einer Vielzahl an Schnittstellen, Diensten und Technologien, wodurch selbst erfahrene Entwickler:innen leicht den Überblick verlieren. Sie könnten IT-Sicherheitsexpert:innen als Überprüfungswerkzeug zum identifizieren von übersehenen Bedrohnungen in der eigenen Analyse dienen, oder das LLM identifiziert die grundlegenden Sicherheitslücken zeiteffizienter und der\*die Experte:in überprüft dessen Output und knüpft daran an, um die komplexeren und kontextspezifischeren Sicherheitsbedrohungen zu finden, die das LLM nicht finden konnte. Denn die Fähigkeit der LLMs, kontextuelle Informationen zu verarbeiten und komplexe Texte zu generieren, macht sie zu potenziell wertvollen Helfern, nicht nur für Privatnutzer:innen, sondern auch für IT-Sicherheitexpert:innen. Dabei Vertrauen, Transparenz und Benutzerfreundlichkeit sind entscheidende Faktoren für einen erfolgreich Einsatz in der Praxis eines solchen Systems.

LLMs werden bereits heute in zahlreichen Anwendungsfeldern genutzt, von der automatisierten Textgenerierung über die Programmierhilfe bis hin zur juristischen oder medizinischen Unterstützung. Auch erste Studien zeigen vielversprechende Ansätze, weisen aber auch auf Herausforderungen hin, z.B. das nicht existente Fehler erkannt werden, Begründungen unklar sind oder fehlender Kontext bei komplexer Systeme zu einer unausreichende Analyse führen. \cite{wu2024threatmodeling}

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.

Warum ist dieses Thema also so relevant?

Zum einen steht die Frage im Raum, wie sich der klassische Threat-Modeling-Prozess durch moderne KI-Tools verbessern lässt, sei es durch Zeitersparnis, durch breitere Abdeckung möglicher Bedrohungen oder durch eine Vereinfachung der Sicherheitsanalyse für nicht-expertische Nutzer:innen. Zum anderen muss geklärt werden, wie vertrauenswürdig die Ergebnisse der KI-Systeme tatsächlich sind: Wie nachvollziehbar sind die Vorschläge? Wo sind ihre Grenzen? Und wie gelingt die Interaktion zwischen Mensch und Maschine in einem sicherheitskritischen Kontext?

Die vorliegende Arbeit beschäftigt sich genau mit diesen Fragen. Sie nimmt eine interdisziplinäre Perspektive ein, indem sie sowohl die technischen Möglichkeiten als auch die praktischen Herausforderungen betrachtet. Dabei geht es nicht nur um die Frage, ob ein LLM wie ChatGPT Bedrohungen erkennen kann, sondern auch darum, wie dieser Prozess gestaltet sein muss, damit er verständlich, nutzbar und vertrauenswürdig ist. Ein besonderer Fokus liegt auf der Zusammenarbeit verschiedener Fachgruppen. Denn Softwareentwicklung, IT-Sicherheit und KI-Technologie sind drei Disziplinen mit unterschiedlichen Perspektiven, Anforderungen und Sprachen. Ein System, das in der einen Domäne als nützlich erachtet wird, kann in der anderen möglicherweise als unbrauchbar angesehen werden, sei es aufgrund von Kontrollmangel oder unzureichenden Anpassungsmöglichkeiten. Deshalb wird in dieser Arbeit, nach Rücksprache mit Experten aus allen drei Domänen, ein Prototyp entwickelt, der ebenfalls getestet und bewertet wird.

Es stellt sich die Frage, wie sich KI-Systeme in sicherheitsrelevante Entscheidungsprozesse integrieren lassen, ohne dass der Mensch die Kontrolle verliert. Dabei muss geregelt sein, wie Risiken kommuniziert werden und unter welchen Bedingungen ein solches System als unterstützend und nicht als bedrohlich empfunden wird.

Die Arbeit versucht den Sicherheitsprozess darauf zu testen mittels LLMs in der Softwareentwicklung effizienter, sicherer und zugänglicher zu machen. Sie verbindet neueste technologische Entwicklungen mit einem klaren gesellschaftlichen Nutzen und leistet dabei einen Beitrag zur verantwortungsvollen Integration von KI in sicherheitskritische Prozesse bei.

**2.Zielsetzung und Leitfrage**

Das Ziel dieser Arbeit ist es zu verstehen, wie Experten für große Sprachmodelle, Softwareentwickler:innen und IT-Sicherheitsexpert:innen mit einem System zur Bedrohungsmodellierung arbeiten, das auf einem LLM 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 LLMs basiert, welches DFDs benutzt, um bei der Bedrohungsmodellierung zu unterstützen. Außerdem wird genauer angeschaut, wie die verschiedenen Fachgruppen diesen Prozess wahrnehmen, welche Erwartungen sie haben und unter welchen Umständen sie ein solches System als hilfreich, vertrauenswürdig und verlässlich einstufen.

In einem ersten Workshop wird in einem Gespräch mit unterschiedlichen Expert:innendiskutiert, 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]](#footnote-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]](#footnote-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.

Mollaeefar und Team (2024)[[3]](#footnote-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]](#footnote-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]](#footnote-5)

 **Ausgangspunkt ist eine aktuelle Studie von Mbaka und Tuma (2024)**, in der untersucht wurde, wie sich DFDs und LLMs auf die Qualität der Bedrohungsbewertung auswirken.  
– Teilnehmende mussten reale und fiktive Bedrohungen in Szenarien wie GitHub-Updates identifizieren.  
– Die Kombination von DFD + LLM (Gruppe C) führte zu den besten Ergebnissen (9,4 korrekte Bewertungen), allerdings auch zu mehr Falsch-Positiven.  
– Weniger erfahrene Nutzer übernahmen häufig unkritisch die Vorschläge des LLM – ein Risiko.

 **Weitere Studien zeigen die Potenziale der Automatisierung:**  
– **Yang et al. (2024)** zeigen, dass LLMs selbstständig DFDs aus Code und Text erzeugen und sogar STRIDE-Bedrohungen klassifizieren können.  
– Das spart Zeit, reduziert Aufwand – ersetzt aber menschliche Expertise nicht vollständig.  
– Das Tool PILLAR (Mollaeefar et al., 2024) geht noch weiter und integriert GDPR & ISO-Normen – hier werden Risiken automatisch erkannt, priorisiert und klassifiziert.

 **Doch DFDs allein sind nicht veraltet**:  
– In traditionellen Ansätzen wie denen von Chis et al. und Scandariato et al. sieht man, dass DFDs + STRIDE immer noch wertvolle Werkzeuge sind – sie schaffen Struktur, fördern systematisches Denken und bleiben lernbar.

 **Die Rolle von Prompt Engineering (Sahoo et al.)**:  
– Die Qualität der Modellantworten hängt stark von der Gestaltung der Prompts ab – durch „Chain of Thought“-Techniken lassen sich bessere Ergebnisse erzielen.  
– Ergänzt durch Adapter-Modelle (Hu et al., 2023), die kleinere, effizientere Modelle ermöglichen, wird KI-basierte Bedrohungsanalyse auch wirtschaftlich interessanter.

 **Zusammenfassend lässt sich sagen:**LLMs revolutionieren das Threat Modeling nicht im Sinne eines Ersatzes, sondern als mächtige Ergänzung.  
Die Kombination klassischer Techniken (DFDs, STRIDE) mit modernen KI-Tools bringt nachweislich bessere Ergebnisse – **aber nur dann**, wenn sie verantwortungsvoll und mit einem geschulten Verständnis angewendet werden.  
Die Zukunft liegt in der **kollaborativen Nutzung von Mensch und Maschine**, mit einem klaren Blick auf die Grenzen beider Seiten.

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. [↑](#footnote-ref-1)
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. [↑](#footnote-ref-2)
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. [↑](#footnote-ref-3)
4. Vgl. R. Scandariato; K. Wuyts; W. Joosen: A descriptive study of Microsoft’s threat modeling technique, am:

   1. Juni 2015. [↑](#footnote-ref-4)
5. Vgl. K. Tuma; W. Mbaka: Human Aspect of Threat Analysis: A Replication, auf: arXiv.org, 2. August 2022. [↑](#footnote-ref-5)