**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.

Ein Vergleich hinsichltich der Frage, was besser funktionert oder ob die LLMs den Menschen erstzen können, wird in dieser Arbeit nicht behandelt. Im Mittelpunkt steht vor allem, die qualitativen Ergebnisse von der Diskussion mit den Expert:innen in ein Prototypen umzuwandeln und auf dieser Basis weitere Erwartungen und Kenntnisse der Proband:innen zu sammeln, die den Prototyp ausprobieren.

Im ersten Workshop wird in einem Gespräch mit unterschiedlichen Expert:innen diskutiert, wie ein solches System effizient gestaltet werden kann. Auf Grundlage ihres Feedbacks, sowie 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.

Die Arbeit erforscht die Erwartungen und das Verständnis der Proband:innen. Dabei wird mit einem expert:innnenzentrierten Design qualtitative Datenerhoben, um die Herausforderung eines KI-gestüttzen Threat-Modeling-Interfaces durch Beispielen zu verdeutlichen und zu vertiefen. Es soll dabei keine spezifische Lösung erarbeitet werden, wie ein „perfektes“ System aussehen könnte. Vielmehr steht ausschließlich die Nutzerperspektive im Mittelpunkt.

**4. Methodik**

Die Studie verfolgt das Ziel, den Einsatz eines LLMs wie ChatGPT 4.0 zur Bedrohungsanalyse auf Basis von DFDs nach dem STRIDE-Modell mit einem nutzerzentrierten Prototyp zu testen. Dazu wird ein mehrstufiges, exploratives methodisches Vorgehen gewählt, um qualitative Erkenntnisse über die technische Umsetzbarkeit sowie die Erwartungshaltung und das Verständnis der Nutzenden zu erforschen.

Zunächst wird ein vorbereiteter DFD als Bild an ChatGPT geschickt. Daraufhin erstellt das LLM Bedrohungsszenarien basierend auf dem STRIDE-Modell. Diese Ergebnisse werden anschließend von Fachleuten aus den Bereichen IT-Sicherheit, Softwareentwicklung und LLM-Technologie geprüft. Das Feedback fließt in die Entwicklung eines Prototyps einer Web-App ein, mit der Nutzer:innen DFDs hochladen und automatische Analysen von ChatGPT erhalten können.

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, wie die Definition von Threat Modeling, STRIDE und DFDs. Außerdem werden das Ziel der Untersuchung und ähnliche wissenschaftliche Arbeiten erläutert. Ein Beispielvideo zeigt, wie ChatGPT 4.0 mit einem DFD „gefüttert” wird und eine Demo-STRIDE-Analyse durchführt. Dieses Video bildet die Grundlage für eine offene Diskussion, in der die Teilnehmenden kritisch hinterfragen, ob die Ergebnisse nachvollziehbar sind, ob wichtige Aspekte übersehen wurden und wie verständlich die Argumentation des Modells ist.

Im Verlauf des Workshops bewerteten die Teilnehmer die vorgestellte Bedrohungsanalyse mit Chat GPT 4.0 hinsichtlich der Qualität der Methode und der praktischen Umsetzbarkeit unter Berücksichtigung ihres Wissens in den Bereichen LLM-Technologie, IT-Sicherheit und Softwareentwicklung. Dabei wurden verschiedene Aspekte nacheinander abgefragt. Zu Beginn wurde eine kurze Einschätzung des Fachwissens der Teilnehmer in IT-Sicherheit abgefragt. Die Teilnehmer schätzten ihr Wissen in den Bereichen IT-Sicherheit und Softwareentwicklung als weitreichend ein, während die LLM-Expertin ihr Wissen in diesem Bereich als geringfügig einschätzte.

Diese Fragen wurden in einem Live-Voting aufgeführt, bei dem die Teilnehmer anonym abstimmen konnten. In diesem Rahmen sprechen wir über die Sinnhaftigkeit des Vorgehens, die spezifischen Herausforderungen der jeweiligen Disziplinen und die Rolle des menschlichen Inputs. Weitere wichtige Themen sind das Vertrauen in KI-gestützte Systeme und die optimale Aufbereitung von Informationen, um die Kommunikation der Nutzer mit dem Tool zu verbessern. Es wurden Fragen zu einem groben Feedback gestellt, ob sich Chat GBT als Threat-Modeling-Hilfstool eignet, was Chat GBT noch an Input oder Output fehlen könnte, welche Erwartungen an ein solches System gestellt werden und ob es überhaupt einen Bedarf in den drei verschiedenen Arbeitsbereichen gibt. Auch konkretere Fragen zu Funktionalität, Usability und Design wurden gestellt, um die Diskussion anzuregen und um eine immer genauere Vorstellung vom Interface zu entwickeln. Gegen Ende wurde auch noch einmal über die Vertrauenswürdigkeit, Transparenz und Nachvollziehbarkeit von LLMs wie Chat GPT diskutiert. Dies ist ein entscheidender Faktor dafür, ob ein Interface dies schaffen kann oder wie ein gewisses Vertrauen in das Tool erzeugt werden kann.

Die Diskussion mit den Expert:innen wird via Audioaufnahme festgehalten, um das Feedback später bearbeiten und auswerten zu können. Jeder Teilnehmer hat vor Beginn des Workshops eine Einwilligungsbescheinigung als Zustimmung zur Audioaufnahme unterschrieben. Die wichtigsten Aspekte aus dem Workshop bzw. der Diskussion werden zusammengefasst und nach weiterer Rücksprache mit den Expert:innen wird ein Prototyp für ein KI-gestütztes Threat-Modeling-Interface entwickelt.

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 achten wir besonders auf die Reaktionen, das Vertrauen in die Technologie, das Interaktionsverhalten und die allgemeine Akzeptanz der Lösung.

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 der aktuellen Forschung zum Threat Modeling lassen sich mehrere Schwerpunkte erkennen. Diese befassen sich mit der Wirksamkeit, Automatisierung und Nutzerfreundlichkeit von Methoden wie STRIDE, DFDs und LLMs. Im Zentrum stehen dabei Fragen nach Genauigkeit und Effizienz sowie nach menschlichen Einflussfaktoren und der Rolle von LLMs in der Bedrohungsidentifikation. Studien zeigen das Potenzial und die Grenzen von LLMs auf. So können LLMs in Kombination mit DFDs die Einschätzung realistischer Bedrohungen verbessern.

Yang et al. (2024) untersuchen die Automatisierung von Threat Modeling mithilfe großer Sprachmodelle (LLMs). Sie zeigen in ihrer Untersuchung „THREATMODELING-LLM“, dass LLMs Datenflussdiagramme (DFDs) aus Systembeschreibungen und Quellcode ableiten, potenzielle Angriffspunkte klassifizieren und Bedrohungsszenarien erstellen können. Es zeigt sich, dass LLMs in der Lage sind, typische Angriffsmethoden zuverlässig zu erkennen und die STRIDE-Kategorien weitgehend korrekt zuzuordnen. Dieser automatisierte Ansatz spart Zeit und reduziert den manuellen Aufwand erheblich, ersetzt jedoch die menschliche Expertise nicht vollständig.

Einschränkungen treten insbesondere bei komplexen oder domänenspezifischen Systemen auf. Laut der Untersuchung ist die Güte der Eingangsdaten und das Kontextwissen des Modells entscheidend für die Qualität der generierten DFDs. Ungenauigkeiten bei der Modellierung, die aus fehlerhaften oder lückenhaften Systemdokumentationen resultieren, haben oft negative Auswirkungen auf die Präzision der Bedrohungsanalyse. Die mangelnde Erklärbarkeit der LLM-Ausgaben bleibt ein entscheidendes Problem, besonders in sicherheitsrelevanten Anwendungen, wo Nachvollziehbarkeit von zentraler Bedeutung ist.

Neben Yang et al. (2024) zeigt auch das Tool PILLAR von Mollaeefar et al. (2024) die Potenziale der Automatisierung im Threat Modeling. PILLAR bindet regulatorische Vorgaben wie die DSGVO und ISO-Normen in die Analyse ein. Dies ermöglicht eine automatische Bewertung der Risiken und eine Priorisierung der festgestellten Bedrohungen. Die beiden Arbeiten machen die Fortschritte, aber auch die Grenzen der LLM-basierten Bedrohungsmodellierung deutlich, vor allem in Bezug auf fachliche Tiefe, Kontextsensitivität und die Notwendigkeit menschlicher Validierung.

Mit PILLAR präsentieren Mollaeefar et al. (2024) ein auf großen LLMs basierendes System zur automatisierten Identifizierung von Datenschutzbedrohungen. Das Tool führt eine automatische Analyse von DFDs durch, erkennt potenzielle Bedrohungen basierend auf dem STRIDE-Modell und bewertet die daraus resultierenden Risiken unter Berücksichtigung bewährter Datenschutz- und Sicherheitsstandards wie der DSGVO (GDPR) und ISO 27001.   
Maschinelles Lernen soll dazu verwendet werden, die Bedrohungen automatisch zu priorisieren. Auf diesem Aspekt liegt ein besonderes Augenmerk. Dadurch kann der Threat-Modeling-Prozess erheblich effizienter gestaltet werden.

Trotz dieser Potenziale erkennt die Studie auch wesentliche Einschränkungen. Hierzu gehören das begrenzte Verständnis des Kontexts seitens LLMs, die eingeschränkte Nachvollziehbarkeit der Entscheidungen der Modelle sowie gelegentlich unangemessene oder falsche Sicherheitsempfehlungen. Außerdem wird das Risiko des Overfittings angesprochen, vor allem wenn das Modell auf spezifischen, nicht generalisierbaren Datensätzen trainiert wurde. PILLAR zeigt vielversprechende Ansätze für die datenschutzkonforme Bedrohungsanalyse mit LLMs, unterstreicht dennoch die Notwendigkeit menschlicher Kontrolle und Kontextwissen.

Scandariato et al. (2015) führten eine der frühesten und umfassendsten Studien zur manuellen Anwendung der STRIDE-Methode durch. 57 Masterstudierende der Informatik führten im Rahmen dieser Untersuchung eine Analyse eines verteilten Softwaresystems durch, um die Effektivität und Praktikabilität von STRIDE in einem realitätsnahen Lehrkontext zu bewerten. Die Ergebnisse zeigten, dass STRIDE grundsätzlich erlern- und anwendbar ist, die Durchführung jedoch mit einem hohen Zeitaufwand verbunden war: Die durchschnittliche Produktivität lag bei lediglich 1,2 bis 1,8 korrekt identifizierten Bedrohungen pro Stunde.

Des Darüber hinaus betrug die durchschnittliche Erkennungsrate weniger als 80 %. Dabei wurden hauptsächlich abstraktere Bedrohungskategorien wie Spoofing oder Repudiation nicht wahrgenommen. Ein weiteres zentrales Ergebnis der Studie war, dass ein höherer Zeitaufwand nicht zwangsläufig zu besseren Analyseergebnissen führte. Ein Hinweis darauf, dass die Effizienzsteigerung durch reine Übung begrenzt ist. Die Studie hebt daher sowohl die Möglichkeiten als auch die Einschränkungen der manuellen Anwendung von STRIDE hervor und bietet wesentliche Hinweise auf den Bedarf an unterstützenden Tools oder Automatisierung.

Tuma und Mbaka (2022) untersuchen in einer theoretisch konzipierten, jedoch nicht umgesetzten Studie, wie menschliche Faktoren wie Geschlecht, kulturelle Herkunft oder Nationalität die Qualität von Bedrohungsanalysen beeinflussen. Auch wenn es keine empirischen Daten gibt, weist die methodische Ausarbeitung darauf hin, dass individuelle Unterschiede in der Wahrnehmung sicherheitskritischer Prozesse erhebliche Auswirkungen auf die Genauigkeit und Objektivität des Threat Modelings haben können. Diese Dimension menschlicher Variabilität lenkt die Aufmerksamkeit auf nicht-technische Einflussfaktoren und betont die Bedeutung interdisziplinärer Perspektiven im Bereich der IT-Sicherheit. Zudem stellt die Studie die Frage, inwieweit standardisierte und automatisierte Analyseprozesse zur Stabilisierung von LLM-basierten Systemen beitragen können. Dennoch emfpiehlt die Arbeit zur Vorsicht im Umgang mit solchen Technologien, insbesondere in Bezug auf potenzielle Bias und das Risiko eines unkritischen Vertrauens in KI-gestützte Entscheidungen. Diese Aspekte wurden in den bestehenden LLM-gestützten Ansätzen bislang kaum berücksichtigt werden.

In einer aktuellen Studie analysieren Mbaka und Tuma (2024) den Einfluss von DFDs und LLMs, insbesondere ChatGPT-3.5 Turbo, auf die Qualität der Bedrohungsanalyse. Im Rahmen eines kontrollierten Experiments wurden Teilnehmende mit technischem Hintergrund aus den Bereichen Softwareentwicklung und Cybersicherheit in vier Gruppen eingeteilt. Diese Gruppen erhielten jeweils verschiedene Arten von Analysehilfen: Eine Gruppe erhielt keine Unterstützung, eine andere ausschließlich DFDs, eine weitere ausschließlich LLM-Unterstützung und die letzte Gruppe eine Kombination aus beiden. Ihre Aufgabe war es, in praxisnahen Szenarien wie GitHub-Updates reale und fiktive Bedrohungen zu erkennen. Die Ergebnisse verdeutlichen, dass die Kombination aus DFDs und LLM (Gruppe C) zu den besten Ergebnissen führte. Die höchste Treffergenauigkeit wurde mit einem Durchschnitt von 9,4 korrekt identifizierten Bedrohungen aus zehn erreicht. Es fiel jedoch gleichzeitig auf, dass diese Gruppe auch eine erhöhte Anzahl an falsch positiven Ergebnissen produzierte. Die Gruppen, die Zugriff auf LLM hatten, erzielten bei der Identifizierung realistischer Bedrohungen signifikant bessere Ergebnisse als die Gruppen ohne technologische Unterstützung. Bei weniger erfahrenen Teilnehmenden offenbarte sich jedoch ein zentrales Risiko: Sie nahmen die Vorschläge des LLM oft ohne kritische Prüfung an. Dies deutet auf eine mögliche Übernahme von Empfehlungen durch Maschinen hin. Die Studie gilt als einer der bisher methodisch differenziertesten empirischen Ansätze zur Evaluation von LLM- und DFD-gestützten Threat-Modeling-Prozessen. Sie macht sowohl das Potenzial für eine Leistungssteigerung durch KI-gestützte Systeme als auch die damit verbundenen Herausforderungen in Bezug auf Nachvollziehbarkeit und menschliches Urteilsvermögen deutlich.

In ihrer Studie aus dem Jahr 2024 heben Sahoo et al. hervor, dass das Prompt Engineering zunehmend Einfluss auf die Qualität von Ausgaben von LLMs hat – vor allem in sicherheitskritischen Bereichen wie der Bedrohungsanalyse. Die Autor:innen ordnen verschiedene Techniken zur Verbesserung von Modellantworten ein und zeigen, dass Methoden wie „Chain of Thought“ (CoT) und „Self-Refine“ insbesondere bei komplexen Aufgaben die Genauigkeit erheblich steigern können. Dadurch ergibt sich für das Threat Modeling ein wesentlicher Vorteil: Die Förderung durch LLMs kann durch eine gezielte Gestaltung der Prompts erheblich verbessert werden, ohne dass die Modelle aufwendig nachtrainiert werden müssen. Die Autor:innen weisen ergänzend auf die Verwendung sogenannter Adapter-Modelle hin (vgl. Hu et al., 2023), die kleinere und effizientere LLM-Varianten ermöglichen. Dadurch werden KI-gestützte Bedrohungsanalysen nicht nur technisch effektiver, sondern auch aus wirtschaftlicher Sicht attraktiver.

Obwohl die Bedeutung herkömmlicher DFDs manchmal angezweifelt wird, belegen Untersuchungen, wie die von Chis et al. (2024), dass DFDs in Verbindung mit bewährten Verfahren wie STRIDE nach wie vor einen hohen praktischen Wert bieten. Sie unterstützen strukturiertes und systematisches Denken, bringen Klarheit in die Analyse und sind leicht zu erlernen. In ihrem designwissenschaftlichen Ansatz schlagen Chis et al. (2024) ein semantisches Modellierungskonzept vor, das klassische DFDs mithilfe von LLMs in maschinenlesbare Wissensgraphen umwandelt. Das Ziel besteht darin, eine frühzeitige Risikoerkennung und eine fundierte Sicherheitsanalyse in komplexen Migrationsszenarien, wie etwa bei der Umstellung auf Cloud-Infrastrukturen, zu ermöglichen. Das Verfahren ermöglicht eine strukturierte und nachvollziehbare Dokumentation von Prozessen, Akteuren und Sicherheitszonen. Es unterstützt Entscheidungsträger:innen bei der Ableitung gezielter Maßnahmen zur Reduzierung von Risiken. Obwohl die Studie noch nicht vollständig veröffentlicht wurde, hebt der Abstract das Potenzial hervor, LLMs bereits in frühen Entwicklungsphasen sinnvoll für Sicherheitsanalysen zu nutzen.

Aufbauend auf diesen Erkenntnissen konkretisieren Chis et al. (2024, BIR Workshops) in einer verwandten Arbeit das Zusammenspiel von STRIDE und DFDs im Kontext von „Secure by Design“. Sie stellen ein Framework vor, das Bedrohungen frühzeitig erkennt, indem es die Secure-by-Design-Prinzipien mit einer systematischen Integration von DFDs und dem STRIDE-Modell verbindet. Am Beispiel eines Online-Shops werden Geschäftsprozesse modelliert und auf sicherheitsrelevante Schwachstellen untersucht. DFDs unterstützen dabei sowohl die Erfassung technischer Risiken, wie beispielsweise Spoofing-Angriffe bei Logins, als auch die Identifizierung nichttechnischer Schwachstellen. Zusammen mit dem STRIDE-Modell erlaubt das Framework eine strukturierte Risikobewertung, zu der auch die Ableitung von Sicherheits-Scores gehört. Diese objektivieren den Analyseprozess und ermöglichen eine fundierte Bewertung potenzieller Bedrohungen. Der Ansatz weist jedoch vor allem bei der Erfassung komplexer und domänenübergreifender Bedrohungsszenarien Schwächen auf. Dennoch betonen die Autor:innen, dass sich mit der modellbasierten Herangehensweise Sicherheitsstrategien systematisch und kontinuierlich verbessern lassen.

**Bereits vor der Verfügbarkeit leistungsstarker LLMs wurde der Automatisierungsgedanke im Threat Modeling verfolgt, wie** die Studie von Singh et al. (2019) zeigt. Sie untersuchten den Einsatz automatisierter Tools auf Basis erweiterter Datenflussdiagramme (eDFDs) im Vergleich zu traditionellen, manuellen Methoden der Bedrohungsanalyse. Der Fokus der Untersuchung lag auf der eSTRIDE-Methode, die klassische DFDs um zusätzliche Metadaten ergänzt und somit eine strukturiertere und automatisierbare Analyse ermöglicht. Zu diesem Zweck wurde ein Prototyp entwickelt, mit dem sich solche eDFDs erstellen und anpassen lassen und der anschließend automatisch Bedrohungen identifiziert. In praktischen Workshops mit Studierenden ohne Vorerfahrung wurde festgestellt, dass die Nutzer:innen des Tools eine deutlich höhere Präzision (74,8 % im Vergleich zu 60 %) und Produktivität (1,4 identifizierte Bedrohungen pro Minute im Vergleich zu 1,2) erzielten als bei manuellen Ansätzen. Der Recall war dabei weitgehend ähnlich. Die Ergebnisse zeigen das Potenzial automatisierter Bedrohungsanalysen zur Effizienzsteigerung, betonen jedoch gleichzeitig die Notwendigkeit einer menschlichen Überprüfung für eine umfassende und qualitativ hochwertige Analyse (Singh et al., 2019).

Zusammenfassend lässt sich aus der aktuellen Studienlage ableiten, dass große Sprachmodelle (LLMs) das Threat Modeling nicht ersetzen, sondern als leistungsfähige Ergänzung zu etablierten Methoden betrachtet werden können. Die Verbindung traditioneller Methoden wie Datenflussdiagrammen (DFDs) und des STRIDE-Modells mit modernen, KI-gestützten Werkzeugen zeigt besonders vielversprechende Ansätze. DFDs bieten eine bewährte visuelle Grundlage für die systematische Analyse von Systemarchitekturen, während LLMs in Kombination mit STRIDE eine automatisierte und beschleunigte Identifikation und Klassifikation potenzieller Bedrohungen ermöglichen.

Forschungen haben nahegelegt, dass diese hybride Herangehensweise die Effizienz und Qualität der Bedrohungsanalyse signifikant steigern kann. LLMs übernehmen dabei Aufgaben wie das automatische Ableiten von Bedrohungsszenarien aus Quellcode oder Systemdokumentationen. Dadurch wird eine zugängliche, natürlichsprachliche Verarbeitung sicherheitsrelevanter Informationen gefördert. Der menschliche Faktor ist jedoch weiterhin von zentraler Bedeutung: Nur durch ein fundiertes Verständnis der verwendeten Modelle, eine kritische Reflexion der Ergebnisse und gezielte Anpassungen an den spezifischen Anwendungskontext kann das volle Potenzial solcher Systeme ausgeschöpft werden.

Zahlreiche Studien weisen jedoch gleichzeitig auf bestehende Herausforderungen hin. Hierzu gehören beispielsweise Gefahren, die aus Overfitting resultieren, aus einer nicht adäquaten Modellierung komplexer oder domänenspezifischer Systeme sowie aus der eingeschränkten Nachvollziehbarkeit bestimmter Ergebnisse von LLM-Generierungen. Das nachhaltige Gelingen automatisierter Bedrohungsanalysen beruht vor diesem Hintergrund vor allem auf der Zusammenarbeit von Mensch und Maschine. Von ausschlaggebender Bedeutung ist, dass die spezifischen Stärken beider Seiten zum Einsatz kommen und ihre Grenzen bedacht berücksichtigt werden.

**Ergebnisse vom Experten-Workshop:**

Zunächst wurde eine grobes Feedback zum einfachen versuch, lediglich Chat GBT mit einem DFD zu füttern und mit mehreren Rückfragen auf ein solides Ergebnis für ein Threat Modeling zu kommen, und da waren sich die Teilnehmer einig, dass dieses Vorgehen nicht ausreiched für ein Hilftool beim Threat Modeling. Im Beispiel wurde sowohl Input Prompts, als auch der Output-Analyse und das miteinhergehende Vertrauen gerade so als ausreichend angesehen. Im Input fehlte ein mehr strukturierter Ansatz, welches beispielsweise das DFD besser erklärt und beschreibt vom Kontext aus.

Die nächsten Fragen fokussierten sich auf das Finden einer passenden Zielgruppe eines solchen KI-gestützen Threat Modeling Tools. Dabei wurde gezeigt, dass ob für Experten in ihrem Arbeitsalltag oder für eine Person ohne Vorwissen in IT-sicherheit, dass es für beide vorstellbar wäre, aber jeweils mit unterschiedlichen Herausforderung und Anwendsungsweise kommt.

Bei Experten wäre es wichtiger, dass das Tool indivdueller auf die Arbeitsbereich abgestimmt ist. Beim allgemeinen Nutzer wäre es wichtiger ein Feature einzubauen, wie eine Prompt-Vorlagen-Bibliothek und, dass die KI auch viel mehr erklärt. Trotzdem wurde hier hervorgehoben, dass ein gewisse Menge an Vorkenntnissen in der IT von Vorteil, wenn nicht sogar Notwendig ist, damit der Nutzer nicht sofort von den vielen Erklärung überwältigt wird. Denn für den normalen Nutzer, wie beispielsweise aus einem kleinen Start-up, wäre ein solches Tool hilfreich, wenn mit einem Experten-Profil als Assistent beim sichern ihrer Systeme den Process mit begleiten könnte.

Eine Chat-Komponente mit der KI sowie das verwenden der Frameworks wie STRIDE, MITRE ATT&CK und PASTA haben alle Teilnehmer als sinnvoll erachtet.

**Audio-Ergebnisse:**

### **Allgemeine Herausforderungen & Überlegungen**

* Chat-GBT ist bei „grüner Wiese“/generischem Input nicht sonderlich hilfreich - es fehlen Kontext und klare Sicherheitsinformationen.
* DFDs sind oft unzureichend, da sie keine implementierten Sicherheitsmaßnahmen enthalten.
* Die Qualität der Analyse steht und fällt mit dem DFD: Magelnder Input, Mangelnder Output.
* Ein reiner One-Shot-Ansatz ist unrealistisch -> Threat Modeling ist ein iterativer, dialogbasierter Prozess.
* Nutzer sollten ermutigt werden, möglichst viele Informationen einzugeben - ChatGPT muss zur Nachfragelogik befähigt sein.
* KI kann nicht vollständig autonom bewerten, es braucht Nutzerfeedback & manuelles Nachbessern.

### **Eingabeformate & Datenquellen**

* Bildformate (z. B. DFD als Grafik) sind schlecht verarbeitbar, textuelle oder strukturierte Formate (z. B. JSON, YAML) sind besser.
* Vorteil textbasierter Formate: günstigere und direktere Verwendung mit der API.
* Möglichkeit, auch den **Output** wieder als JSON zu bekommen, um ihn weiterzuverarbeiten oder zu kürzen.
* Nummerierung von DFD-Kanten erlaubt gezielte Rückmeldung zu einzelnen Datenflüssen.

### **Zielgruppen & Benutzerprofile**

* Für Nicht-Security-Experten ist die Tiefe der Ausgabe oft zu viel, es braucht didaktische Aufbereitung.
* Entwickler interessieren sich mehr für Frameworks, Bibliotheken und technische Umsetzungstipps.
* Kleine Firmen/Freelancer profitieren besonders, da oft Know-how fehlt oder später nachgeholt wird.
* Security-Profis würden das System eher als Zweitgutachter oder Challenge-Partner nutzen.
* System sollte persönliche Expertenprofile berücksichtigen: je nach Vorwissen differenzierte Erklärungen anbieten.

### **Interaktion & Nutzerführung**

* Das System sollte interaktiv Feedback einholen: **„Hier fehlen mir Infos, bitte nachspezifizieren“**.
* Threat Modeling sollte eingebettet sein in agiles oder iteratives Vorgehen (Spiralmodell).
* Das Tool sollte auch Systeme/ Ansätze von vor Wochen analysieren können, was hat sich seit letzter Iteration geändert?
* System soll helfen, das DFD gemeinsam zu verbessern, nicht nur auswerten, sondern mitentwickeln.
* Das Interface sollte den Nutzer stärker anleiten: „Was fehlt noch?“, „Was wurde schon implementiert?“.
* Automatisches Parsen der Antwort nötig: nicht rohe ChatGPT-Ausgabe zeigen, sondern strukturieren.

### **Prompt-Design & Automation**

* Prompt-Vorlagen könnten helfen, müssen aber sorgfältig genutzt werden (Gefahr: Anchoring-Effekt).
* Vorlagen können helfen, Themen zu erschließen, in denen der Nutzer kein Experte ist (z. B. Krypto).
* Gefahr der Überautomatisierung: Nutzer verfallen in „Autopilot“, vertrauen dem System zu sehr.
* Hintergrund-Prompts sind sinnvoll, sollten aber die Eigenverantwortung des Nutzers nicht verdrängen.
* Vorschlag: Prompts und Ergebnisse intern validieren oder mit bekannten Mustern/Datenbanken kombinieren.

### **Transparenz & Vertrauenswürdigkeit**

* System soll seine **Unsicherheiten** klar kommunizieren („Ich vermute, weil ich es nicht genau sehe...“).
* Vertrauen steigt mit menschlicher, transparenter Kommunikation: wo liegen Unsicherheiten?
* Persönliche Erklärungstiefe: Experten brauchen keine Basisinfos, Personen ohne Fachwissen hingegen schon.
* Wichtiger Hinweis: LLMs bleiben fehleranfällig -> Nutzer müssen darauf hingewiesen werden.

### **Zielbild für das Tool**

* Ein **interaktives System**, das Nutzer durch die Modellierung begleitet und bei der Verbesserung unterstützt.
* Fokus auf **Bedrohungen erkennen**, nicht unbedingt gleich Lösungen liefern.
* Ideal für Architekten als Checkliste/Review-Tool zur Überprüfung, ob relevante Threats bedacht wurden.
* Nicht Ziel: Komplettlösung für Security-Design, sondern **Assistenzsystem im Prozess**.