LLM AI Cybersecurity & Governance Checklist

From the OWASP Top 10
for LLM Applications Team

Version: 1.0

Published: *February 19, 2024*

Revision History

Revision	Date	Author(s)	Description
0.1	2023-11-01	Sandy Dunn	initial draft
0.5	2023-12-06	SD, Team	public draft
0.9	2023-02-15	SD, Team	pre-release draft
1.0	2024-02-19	SD, Team	public release v 1.0

Die in diesem Dokument bereitgestellten Informationen stellen keine rechtliche Beratung dar und sind auch nicht als solche gedacht. Alle Informationen dienen nur allgemeinen Informationszwecken.

Dieses Dokument enthält Links zu anderen Websites von Dritten. Solche Links dienen nur zur Bequemlichkeit und OWASP empfiehlt oder unterstützt nicht den Inhalt der Websites von Dritten.

Überblick

Jeder Internetnutzer und jedes Unternehmen sollte sich auf die bevorstehende Welle leistungsstarker generativer künstlicher Intelligenz (GenAI) Anwendungen vorbereiten. GenAI hat enormes Potenzial für Innovation, Effizienz und kommerziellen Erfolg in einer Vielzahl von Branchen. Dennoch bringt es wie jede leistungsstarke Technologie in der frühen Entwicklungsphase eigene offensichtliche und unerwartete Herausforderungen mit sich.

Künstliche Intelligenz hat sich in den letzten 50 Jahren stark weiterentwickelt und unterstützt unauffällig eine Vielzahl von Unternehmensprozessen, bis das Auftreten von ChatGPT die Entwicklung und Nutzung von Large Language Models (LLMs) sowohl bei Einzelpersonen als auch in Unternehmen vorangetrieben hat. Anfangs waren diese Technologien auf den akademischen Bereich oder die Durchführung bestimmter, aber wichtiger Aktivitäten innerhalb von Unternehmen beschränkt, die nur einer ausgewählten Gruppe zugänglich waren. Doch durch Fortschritte bei der Datenverfügbarkeit, Rechenleistung, GenAI-Fähigkeiten und die Veröffentlichung von Tools wie Llama 2, ElevenLabs und Midjourney ist KI von einer Nische zu einer allgemein weit verbreiteten Akzeptanz geworden. Diese Verbesserungen haben GenAI-Technologien nicht nur zugänglicher gemacht, sondern auch deutlich gemacht, dass Unternehmen solide Strategien für die Integration und Nutzung von KI in ihren Betrieb entwickeln müssen. Dies stellt einen großen Fortschritt dar in der Art und Weise, wie wir Technologie nutzen.

- **Künstliche Intelligenz** (KI) ist ein weiter Begriff, der alle Bereiche der Informatik umfasst, die es Maschinen ermöglichen, Aufgaben zu erledigen, die normalerweise menschliche Intelligenz erfordern würden. Maschinelles Lernen und generative KI sind zwei Unterkategorien von KI.

- **Maschinelles Lernen** ist eine Teilmenge von KI, die sich auf die Entwicklung von Algorithmen konzentriert, die aus Daten lernen können. Maschinelles Lernen Algorithmen werden anhand eines Datensatzes trainiert und können dann diese Daten verwenden, um Vorhersagen oder Entscheidungen über neue Daten zu treffen.

- **Generative KI** ist eine Art des maschinellen Lernens, die sich darauf konzentriert, neue Daten zu erstellen.

- Ein **Large Language Model (LLM)** ist ein Typ von KI-Modell, das menschenähnlichen Text verarbeitet und generiert. Im Kontext der künstlichen Intelligenz bezieht sich ein "Modell" auf ein System, das darauf trainiert ist, Vorhersagen basierend auf Eingabedaten zu treffen. LLMs werden speziell auf großen Datensätzen natürlicher Sprache trainiert und tragen daher den Namen Large Language Models.

Organisationen betreten Neuland bei der Sicherung und Überwachung von GenAI-Lösungen. Der schnelle Fortschritt von GenAI eröffnet auch Angreifern Möglichkeiten, ihre Angriffsstrategien zu verbessern und stellt somit eine doppelte Herausforderung in Verteidigung und Bedrohung dar.

Unternehmen nutzen künstliche Intelligenz in vielen Bereichen, darunter Personalbeschaffung in der Personalabteilung, E-Mail-Spam-Filterung, Verhaltensanalyse bei SIEM und verwaltete Detection- und Response-Anwendungen. Der Schwerpunkt dieses Dokuments liegt jedoch hauptsächlich auf Large Language Model-Anwendungen und ihrer Funktion bei der Erzeugung von generierten Inhalten.

Verantwortungsvolle und vertrauenswürdige künstliche Intelligenz

Mit dem Aufkommen von Herausforderungen und Vorteilen der künstlichen Intelligenz - und mit der Verabschiedung von Vorschriften und Gesetzen - entwickeln sich die Prinzipien und Grundpfeiler für die verantwortungsvolle und vertrauenswürdige Nutzung von KI von idealistischen Überlegungen zu etablierten Standards. Die OWASP AI Exchange Working Group überwacht diese Veränderungen und behandelt die breiteren und anspruchsvolleren Überlegungen für alle Aspekte der künstlichen Intelligenz.

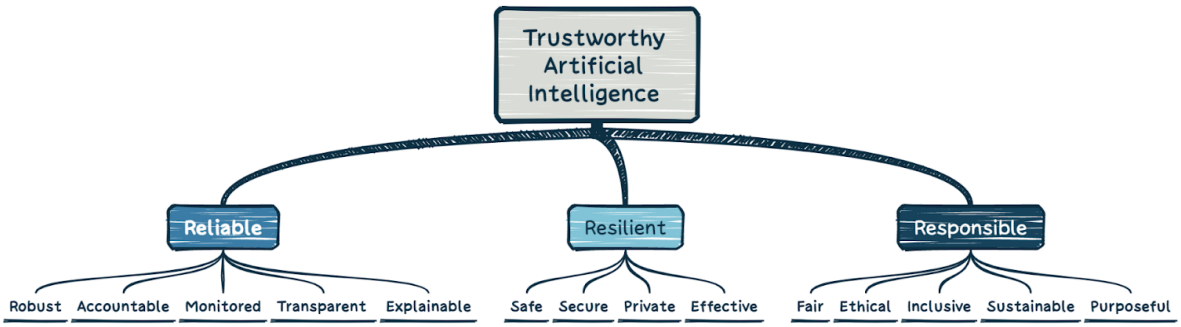

Figure 1.1: Bild, das die Grundpfeiler der vertrauenswürdigen künstlichen Intelligenz darstellt

An wen richtet sich dieses Dokument?

Die OWASP Top 10 für LLM-Anwendungen Cybersecurity- und Governance-Checkliste richtet sich an Führungskräfte in den Bereichen Business, Technologie, Cybersicherheit, Datenschutz, Compliance und Recht, DevSecOps, MLSecOps sowie an Cybersicherheitsteams und Verteidiger. Sie richtet sich an Personen, die bestrebt sind, in der schnelllebigen AI-Welt voraus zu sein, um nicht nur KI für den Unternehmenserfolg zu nutzen, sondern auch um sich gegen die Risiken hastiger oder unsicherer KI-Implementierungen zu schützen. Diese Führungskräfte und Teams müssen Taktiken entwickeln, um Chancen zu nutzen, Herausforderungen zu bewältigen und Risiken zu mindern.

Diese Checkliste soll diesen Technologie- und Business-Führungskräften helfen, die Risiken und Vorteile der Verwendung von LLM schnell zu verstehen und ihnen ermöglichen, sich auf die Entwicklung einer umfassenden Liste kritischer Bereiche und Aufgaben zu konzentrieren, die erforderlich sind, um die Organisation zu verteidigen und zu schützen, während sie eine Large Language Model-Strategie entwickeln.

Das Team der OWASP Top 10 für LLM-Anwendungen hofft, dass diese Liste Organisationen dabei helfen wird, ihre bestehenden Verteidigungstechniken zu verbessern und Techniken zu entwickeln, um mit den neuen Bedrohungen umzugehen, die sich aus der Verwendung dieser aufregenden Technologie ergeben.

Warum eine Checkliste?

Checklisten zur Formulierung von Strategien verbessern die Genauigkeit, definieren Ziele, erhalten Einheitlichkeit und fördern gezielte, bewusste Arbeit, wodurch Fehler und übersehene Details reduziert werden. Das Befolgen einer Checkliste erhöht nicht nur das Vertrauen in eine sichere Einführungsreise, sondern fördert auch zukünftige Innovationsmöglichkeiten von Organisationen, indem sie eine einfache und effektive Strategie für kontinuierliche Verbesserung bereitstellt.

Nicht umfassend

Obwohl dieses Dokument Organisationen bei der Entwicklung einer LLM-Strategie in einer sich schnell verändernden technischen, rechtlichen und regulatorischen Umgebung unterstützen soll, ist es nicht umfassend und deckt nicht jeden Anwendungsfall oder jede Verpflichtung ab. Bei der Verwendung dieses Dokuments sollten Organisationen Assessments und Praktiken über den Umfang der bereitgestellten Checkliste hinaus erweitern, wie es für ihren Anwendungsfall oder ihre Rechtsprechung erforderlich ist.

Herausforderungen bei Large Language Models

Large Language Models stehen vor mehreren ernsthaften und einzigartigen Herausforderungen. Eine der wichtigsten ist, dass bei der Arbeit mit LLMs die Kontroll- und Datenflächen nicht streng isoliert oder trennbar sein können. Eine weitere bedeutende Herausforderung besteht darin, dass LLMs von Natur aus nichtdeterministisch sind und unterschiedliche Ergebnisse liefern, wenn sie aufgefordert oder angefordert werden. LLMs verwenden semantische Suche anstelle von Schlüsselwortsuche. Der wesentliche Unterschied zwischen den beiden besteht darin, dass der Algorithmus des Modells die Begriffe in seiner Antwort priorisiert. Dies ist eine bedeutende Abweichung davon, wie Verbraucher zuvor Technologie genutzt haben, und hat Auswirkungen auf die Konsistenz und Zuverlässigkeit der Ergebnisse. Halluzinationen, die aus den Lücken und Schulungsmängeln in den Daten des Modells

entstehen, sind das Ergebnis dieser Methode.

Es gibt Methoden, um die Zuverlässigkeit zu verbessern und die Angriffsfläche für Jailbreaking, Modelltrickserei und Halluzinationen zu verringern, aber es besteht ein Kompromiss zwischen Einschränkungen und Nützlichkeit in Bezug auf Kosten und Funktionalität.

Die Verwendung von LLM und LLM-Anwendungen erhöht die Angriffsfläche einer Organisation. Einige mit LLMs verbundene Risiken sind einzigartig, aber viele sind bekannte Probleme, wie beispielsweise die bekannte Software-Bestandsliste (SBoM), Lieferketten, Datenschutzverlustschutz (DLP) und autorisierter Zugriff. Es gibt auch erhöhte Risiken, die nicht direkt mit GenAI zusammenhängen, aber GenAI die Effizienz, Fähigkeit und Effektivität von Angreifern erhöht, die Organisationen angreifen und bedrohen.

Angreifer nutzen zunehmend LLM- und generative KI-Tools, um herkömmliche Methoden zur Angriff auf Unternehmen, Einzelpersonen und Regierungssysteme zu verbessern und zu beschleunigen. LLM erleichtert ihre Fähigkeit, Techniken zu verbessern, die es ihnen ermöglichen, mühelos neue Malware zu erstellen, die möglicherweise neuartige Zero-Day-Schwachstellen enthält oder darauf abzielt, erkannt zu werden. Sie können auch anspruchsvolle, einzigartige oder maßgeschneiderte Phishing-Schemata generieren. Die Erstellung überzeugender Deepfakes, ob Video oder Audio, fördert weiterhin ihre Social Engineering-Strategien. Darüber hinaus ermöglichen ihnen diese Tools, Eindringungen auszuführen und innovative Hacking-Fähigkeiten zu entwickeln. In Zukunft wird der "maßgeschneiderte" und kombinierte Einsatz von KI-Technologie durch kriminelle Akteure spezifische Antworten und dedizierte Lösungen für die angemessene Verteidigung und Resilienz einer Organisation erfordern.

Organisationen stehen auch der Bedrohung gegenüber, die Fähigkeiten von LLMs nicht zu nutzen, beispielsweise durch einen Wettbewerbsnachteil, eine Marktwahrnehmung durch Kunden und Partner als veraltet, die Unfähigkeit, personalisierte Kommunikation im großen Maßstab zu skalieren, Innovationsstagnation, operative Ineffizienzen, ein höheres Risiko menschlicher Fehler in Prozessen und ineffiziente Zuweisung von menschlichen Ressourcen.

Das Verständnis der verschiedenen Arten von Bedrohungen und deren Integration in die Geschäftsstrategie wird dazu beitragen, sowohl die Vor- als auch die Nachteile der Verwendung von Large Language Models (LLMs) gegenüber der Nichtverwendung abzuwägen und sicherzustellen, dass sie die Erfüllung der Geschäftsziele des Unternehmens beschleunigen und nicht behindern.

LLM-Bedrohungskategorien

Figure 1.2: Bild, das die Arten von KI-Bedrohungen darstellt

Schulungen zur Sicherheit und Privatsphäre von Künstlicher Intelligenz

Mitarbeiter in Organisationen profitieren von Schulungen, um künstliche Intelligenz, generative künstliche Intelligenz und die zukünftigen potenziellen Folgen des Aufbaus, Kaufs oder der Nutzung von LLMs zu verstehen. Schulungen für die zulässige Nutzung und Sicherheitsbewusstsein sollten sich an alle Mitarbeiter richten und für bestimmte Positionen wie Personalwesen, Recht, Entwickler, Datenteams und Sicherheitsteams spezialisierter sein.

Richtlinien für faire Nutzung und gesunde Interaktion sind wichtige Aspekte, die von Anfang an eingebunden werden sollten und ein Eckpfeiler für den Erfolg zukünftiger Sensibilisierungskampagnen zur Cybersicherheit von KI sind. Dadurch erhalten Benutzer das Wissen über die grundlegenden Regeln für die Interaktion sowie die Fähigkeit, gutes Verhalten von schlechtem oder unethischem Verhalten zu unterscheiden.

Integrieren von LLM-Sicherheit und Governance in bestehende etablierte Praktiken und Kontrollen

Obwohl KI und generative KI eine neue Dimension in Bezug auf Cybersicherheit, Resilienz, Datenschutz und die Erfüllung gesetzlicher und regulatorischer Anforderungen darstellen, sind die bewährten Praktiken, die seit langem existieren, immer noch der beste Weg, um Probleme zu identifizieren, Schwachstellen zu finden, sie zu beheben und potenzielle Sicherheitsprobleme zu mindern.

- Stellen Sie sicher, dass das Management von KI-Systemen in bestehende organisatorische Praktiken integriert ist.

- Stellen Sie sicher, dass AIML-Systeme bestehenden Datenschutz-, Governance- und Sicherheitspraktiken folgen, wobei spezifische Datenschutz-, Governance- und Sicherheitspraktiken für KI implementiert werden, wenn erforderlich.

Grundlegende Sicherheitsprinzipien

LLM-Fähigkeiten bringen eine andere Art von Angriff und Angriffsfläche mit sich. LLMs sind anfällig für komplexe Geschäftslogikfehler wie Prompt-Injection, unsicheres Plugin-Design und Remote Code Execution. Bestehende bewährte Praktiken sind der beste Weg, um diese Probleme zu lösen. Ein internes Produktsicherheitsteam, das sich mit sicherer Softwareüberprüfung, Architektur, Datenverwaltung und Bewertungen von Drittanbietern auskennt, sollte überprüfen, wie stark die aktuellen Kontrollen sind, um Probleme zu finden, die durch LLM wie Stimmerzeugung, Imitation oder Umgehen von Captchas verschlimmert werden könnten. Angesichts der jüngsten Fortschritte im maschinellen Lernen, in der NLP (Natural Language Processing), NLU (Natural Language Understanding), Deep Learning und in jüngster Zeit in LLMs (Large Language Models) und generativer KI wird empfohlen, Fachleute, die in diesen Bereichen versiert sind, neben Cybersicherheits- und DevOps-Teams einzusetzen. Ihre Expertise wird nicht nur bei der Einführung dieser Technologien helfen, sondern auch bei der Entwicklung innovativer Analysen und Reaktionen auf aufkommende Herausforderungen.

Risiko

Der Bezug auf Risiko verwendet die Definition nach ISO 31000: Risiko = "Auswirkung von Unsicherheit auf Ziele". Die in der Checkliste enthaltenen LLM-Risiken umfassen eine gezielte Liste von LLM-Risiken, die adversäre, sicherheitsrelevante, rechtliche, regulatorische, reputationsbezogene, finanzielle und wettbewerbsbezogene Risiken behandeln.

Vulnerability- Und Mitigations-Taxonomie

Aktuelle Systeme zur Klassifizierung von Schwachstellen und zum Austausch von Bedrohungsinformationen, wie OVAL, STIX, CVE und CWE, entwickeln noch die Fähigkeit, Verteidiger über Schwachstellen und Bedrohungen, die spezifisch für Large Language Models (LLMs) und Predictive Models sind, zu überwachen und zu alarmieren. Es wird erwartet, dass Organisationen sich auf diese etablierten und anerkannten Standards wie CVE zur Klassifizierung von Schwachstellen und STIX zum Austausch von Cybersicherheitsbedrohungsdaten (CTI) verlassen werden, wenn Schwachstellen oder Bedrohungen für KI/ML-Systeme und ihre Lieferketten identifiziert werden.

Entwickeln der LLM-Strategie

Die rasante Expansion von Large Language Model (LLM)-Anwendungen hat die Aufmerksamkeit und Untersuchung aller KI/ML-Systeme erhöht, die in Geschäftsprozessen eingesetzt werden, einschließlich sowohl Generative AI als auch etablierter Predictive AI/ML-Systeme. Diese verstärkte Aufmerksamkeit macht potenzielle Risiken sichtbar, wie zum Beispiel Angriffe auf Systeme, die zuvor übersehen wurden, sowie Governance- oder rechtliche Herausforderungen, die in Bezug auf rechtliche, Datenschutz-, Haftungs- oder Garantiefragen vernachlässigt wurden. Für jede Organisation, die KI/ML-Systeme in ihren Geschäftsprozessen einsetzt, ist es entscheidend, umfassende Richtlinien, Governance-Maßnahmen, Sicherheitsprotokolle, Datenschutzmaßnahmen und Rechenschaftsstandards zu bewerten und festzulegen, um sicherzustellen, dass diese Technologien sicher und ethisch mit den Geschäftsprozessen übereinstimmen.

Angreifer oder Gegner stellen die unmittelbarste und schädlichste Bedrohung für Unternehmen, Personen und Regierungsbehörden dar. Ihre Ziele, die von finanziellen Gewinnen bis hin zu Spionage reichen, treiben sie dazu an, kritische Informationen zu stehlen, den Betrieb zu stören und das Vertrauen zu schädigen. Darüber hinaus erhöht ihre Fähigkeit, neue Technologien wie KI und maschinelles Lernen einzusetzen, die Geschwindigkeit und Raffinesse von Angriffen, so dass es für Abwehrmaßnahmen schwierig ist, den Angriffen voraus zu sein.

Die dringlichste nicht-gegnerische LLM-Bedrohung für viele Organisationen geht von "Schatten-AI" aus: Mitarbeiter verwenden nicht genehmigte Online-KI-Tools, unsichere Browser-Plugins und Drittanbieter-Anwendungen, die LLM-Funktionen über Updates oder Upgrades einführen und damit die standardmäßigen Softwaregenehmigungsprozesse umgehen.

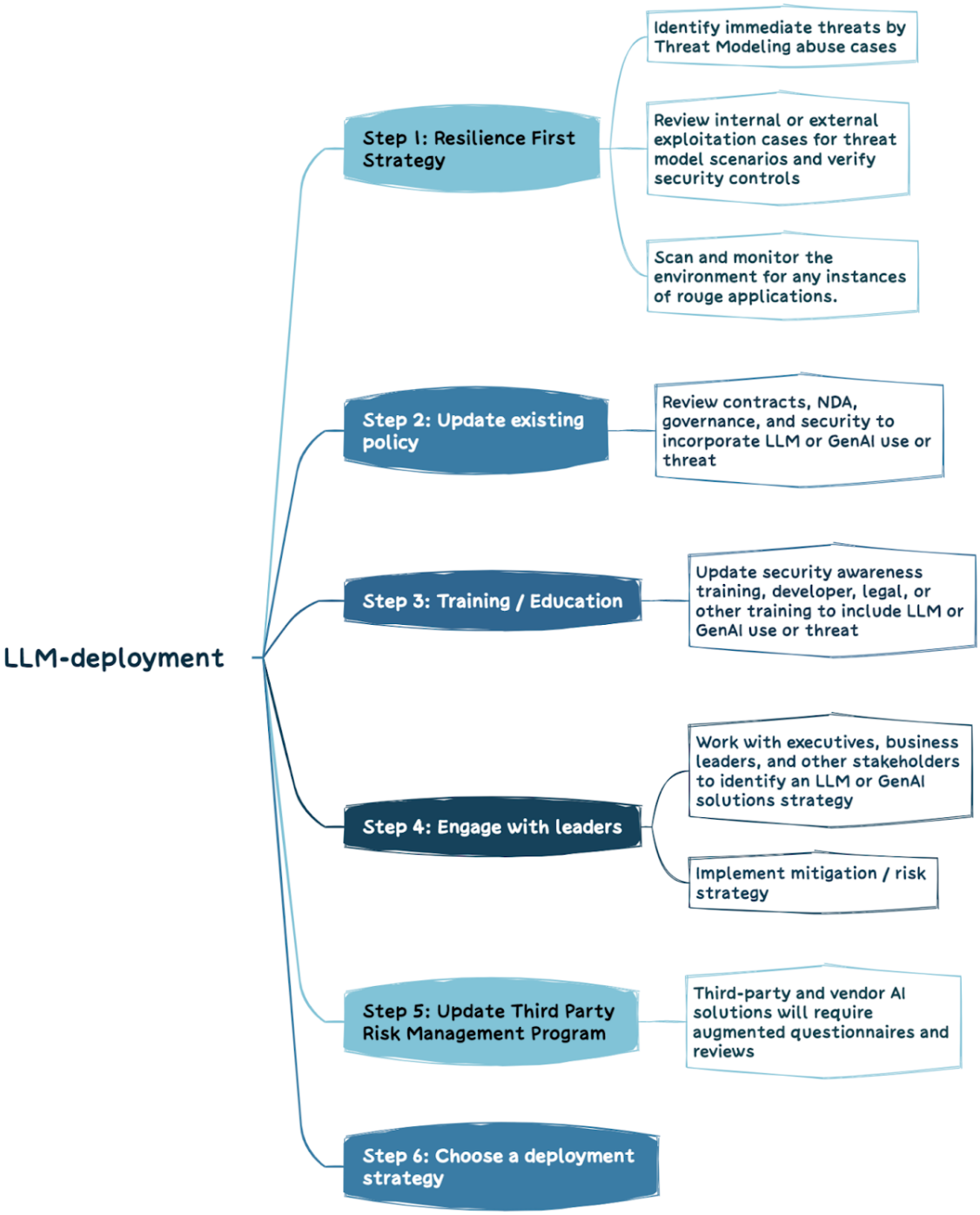

Figure 2.1: Bild der Optionen für die Deployment-Strategie

Deployment-Strategie

Die Bereiche reichen von der Nutzung öffentlicher Verbraucheranwendungen bis hin zum Training proprietärer Modelle auf privaten Daten. Faktoren wie die Empfindlichkeit des Anwendungsfalls, benötigte Fähigkeiten und verfügbare Ressourcen helfen dabei, das richtige Gleichgewicht zwischen Bequemlichkeit und Kontrolle zu finden. Das Verständnis dieser fünf Modelltypen bietet jedoch einen Rahmen zur Bewertung der Optionen.

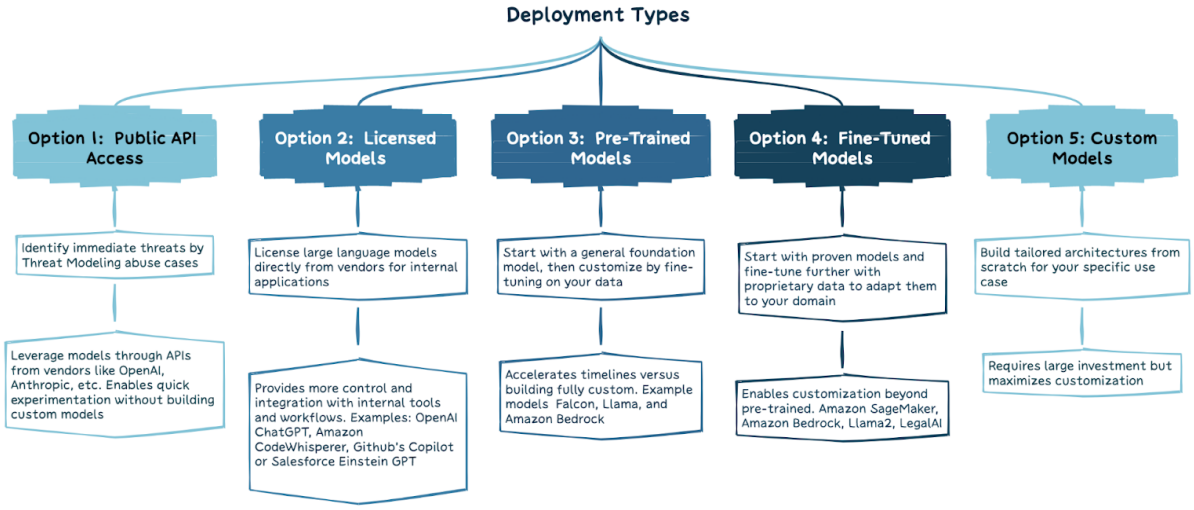

Figure 2.2: Bild der Optionen für die Deployment-Typen

Checkliste

Risiko durch Gegner(Adversial Risk)

Risiko durch Gegner umfasst Konkurrenten und Angreifer.

- ☐ Untersuchen Sie, wie Konkurrenten in künstliche Intelligenz investieren. Obwohl es Risiken bei der Einführung von KI gibt, gibt es auch Geschäftsvorteile Vorteile, die sich auf zukünftige Marktpositionen auswirken können.
- ☐ Untersuchen Sie den Einfluss bestehender Kontrollen wie Passwort-Resets auf Basis von Sprache, die möglicherweise keinen angemessenen defensiven Schutz vor neuen GenAI-gesteigerten Angriffen mehr bieten.
- ☐ Aktualisieren Sie Ihren Incident-Response-Plan und die Playbooks für GenAI-gesteigerte Angriffe und AIML-spezifische Vorfälle.

Bedrohungsmodellierung

Eine Bedrohungsmodellierung (Threat modeling) wird dringend empfohlen, um Bedrohungen zu identifizieren und Prozesse und Sicherheitsvorkehrungen zu untersuchen. Die Bedrohungsmodellierung ist ein Satz systematischer, wiederholbarer Prozesse, die es ermöglichen, vernünftige Sicherheitsentscheidungen für Anwendungen, Software und Systeme zu treffen. Die Bedrohungsmodellierung für GenAI-beschleunigte Angriffe und vor der Implementierung von LLMs ist der kostengünstigste Weg, um Risiken zu identifizieren und zu mindern, Daten zu schützen, die Privatsphäre zu gewährleisten und eine sichere, konforme Integration in Ihr Unternehmen sicherzustellen.

- □ Wie werden Angreifer Angriffe gegen die Organisation, Mitarbeiter, Führungskräfte oder Benutzer beschleunigen? Organisationen sollten "hyperpersonalisierte" Angriffe in großem Maßstab unter Verwendung von Generative AI antizipieren. LLM-unterstützte Spear-Phishing-Angriffe sind nun exponentiell effektiver, gezielter und für einen Angriff gewappnet.
- □ Wie könnte GenAI für Angriffe auf Kunden oder Klienten des Unternehmens verwendet werden, z.B. durch Spoofing oder GenAI-generierten Inhalt?
- □ Kann das Unternehmen schädliche oder bösartige Eingaben oder Anfragen an LLM-Lösungen erkennen und neutralisieren?
- □ Kann das Unternehmen Verbindungen zu bestehenden Systemen und Datenbanken mit sicheren Integrationen an allen LLM-Vertrauensgrenzen schützen?
- □ Verfügt das Unternehmen über Maßnahmen zur Verhinderung von internen Bedrohungen, um Missbrauch durch autorisierte Benutzer zu verhindern?
- □ Kann das Unternehmen unbefugten Zugriff auf proprietäre Modelle oder Daten verhindern, um das geistige Eigentum zu schützen?
- □ Kann das Unternehmen die Erzeugung von schädlichem oder unangemessenem Inhalt mit automatisierter Inhaltsfilterung verhindern?

AI-Asset-Inventar

Ein AI-Asset-Inventar sollte sowohl für intern entwickelte als auch externe oder Drittanbieterlösungen gelten.

- □ Katalogisieren Sie vorhandene KI-Dienste, -Tools und -Besitzer. Kennzeichnen Sie spezifische Inventarpositionen im Asset-Management.
- □ Fügen Sie KI-Komponenten zur Software-Bill-of-Materials (SBOM) hinzu, einer umfassenden Liste aller Softwarekomponenten, Abhängigkeiten und Metadaten, die mit Anwendungen verbunden sind.
- □ Erfassen Sie KI-Datenquellen und die Sensitivität der Daten (geschützt, vertraulich, öffentlich)
- □ Stellen Sie fest, ob Penetrationstests oder Red Teaming von implementierten KI-Lösungen erforderlich sind, um das derzeitige Angriffsflächenrisiko zu bestimmen.
- □ Erstellen Sie einen AI-Lösungs-Onboarding-Prozess.
- □ Stellen Sie sicher, dass qualifiziertes IT-Administrationspersonal intern oder extern vorhanden ist und den Anforderungen der SBOM entspricht.

AI-Sicherheits- und Datenschutzschulung

- □ Sprechen Sie sich aktiv mit Ihren Mitarbeitern, um Bedenken im Zusammenhang mit geplanten LLM-Initiativen zu verstehen und anzugehen.
- □ Schaffen Sie eine Kultur offener und transparenter Kommunikation über den Einsatz von vorhersagender oder generativer AI im organisatorischen Prozess, den Systemen, der Mitarbeiterverwaltung und -unterstützung sowie den Kundeninteraktionen und wie deren Einsatz geregelt, verwaltet und Risiken angegangen werden.
- □ Schulen Sie alle Benutzer in Ethik, Verantwortung und rechtlichen Fragen wie Gewährleistung, Lizenzen und Urheberrecht.
- □ Aktualisieren Sie das Security-Awareness-Training, um GenAI-bezogene Bedrohungen einzubeziehen. Sprach- und Bildklonung sowie anbetrachts zunehmender Spear-Phishing-Angriffe.
- □ Jede übernommene GenAI-Lösung sollte Schulungen für DevOps und Cybersicherheit für den Deployment-Prozess einschließen, um die Sicherheit und Vertrauenswürdigkeit von KI zu gewährleisten.

Business-Cases festlegen

Solide Business-Cases sind entscheidend für die Bestimmung des Geschäftswerts einer vorgeschlagenen KI-Lösung, um Risiken und Vorteile abzuwägen und die Rentabilität zu bewerten und zu testen. Es gibt eine enorm große Anzahl von potenziellen Anwendungsfällen, von denen einige Beispiele gegeben sind.

- □ Verbesserung der Kundenerfahrung
- □ Bessere operative Effizienz
- □ Besseres Wissensmanagement
- □ Verbesserte Innovation
- □ Marktforschung und Wettbewerbsanalyse
- □ Dokumentenerstellung, Übersetzung, Zusammenfassung und Analyse

Governance

Governance für LLMs ist erforderlich, um Organisationen Transparenz und Rechenschaftspflicht zu bieten. Die Identifizierung von AI-Plattform- oder Prozessbesitzern, die möglicherweise mit der Technologie oder den ausgewählten Anwendungsfällen des Unternehmens vertraut sind, ist nicht nur ratsam, sondern auch notwendig, um eine angemessene Reaktionsgeschwindigkeit sicherzustellen, die Schäden an etablierten Unternehmensdigitalprozessen verhindert.

- □ Erstellen Sie die KI-RACI-Tabelle der Organisation (wer ist verantwortlich, wer ist zuständig, wer sollte konsultiert werden und wer sollte informiert werden)
- □ Dokumentieren und zuweisen von AI-Risiken, Risikobewertungen und Governance-Verantwortung innerhalb der Organisation.
- □ Erstellen Sie Richtlinien für das Datenmanagement, einschließlich technischer Durchsetzung, in Bezug auf die Klassifizierung von Daten und die Beschränkungen bei der Verwendung. Modelle sollten nur Daten verwenden, die für das Mindestzugriffsniveau eines Benutzers des Systems klassifiziert sind. Aktualisieren Sie beispielsweise die Datenschutzrichtlinie, um zu betonen, dass keine geschützten oder vertraulichen Daten in nicht unternehmensverwaltete Tools eingegeben werden dürfen.
- □ Erstellen Sie eine KI-Richtlinie, die von etablierten Richtlinien unterstützt wird (z.B. Standard für gutes Verhalten, Datenschutz, Softwareverwendung)
- □ Veröffentlichen Sie eine akzeptable Verwendungsmatrix für verschiedene generative KI-Tools, die von Mitarbeitern verwendet werden können.
- □ Dokumentieren Sie die Quellen und Verwaltung aller Daten, die die Organisation aus den generativen LLM-Modellen verwendet.

Rechtliche Aspekte

Viele der rechtlichen Implikationen von KI sind undefiniert und potenziell sehr kostspielig. Eine Partnerschaft zwischen IT, Sicherheit und Recht ist entscheidend, um Lücken zu identifizieren und unklare Entscheidungen anzugehen.

- □ Bestätigen Sie, dass Produkthaftungsgarantien im Produktentwicklungsprozess klar sind, um festzulegen, wer für Produkthaftung mit KI verantwortlich ist.
- □ Überprüfen und aktualisieren Sie bestehende Geschäftsbedingungen für etwaige GenAI-Aspekte.
- □ Überprüfen Sie Vereinbarungen über Endbenutzerlizenzvereinbarungen (EULA) für GenAI-Plattformen. EULA für GenAI-Plattformen unterscheiden sich stark in der Art und Weise, wie sie Benutzerhinweise, Ausgaberechte und -eigentum, Datenschutz, Compliance, Haftung, Privatsphäre und Einschränkungen bei der Verwendung von Ausgaben behandeln.
- □ Ändern Sie Endbenutzervereinbarungen für Kunden, um das Unternehmen von Haftungsrisiken im Zusammenhang mit Plagiat, Verbreitung von Voreingenommenheit oder Verletzung des geistigen Eigentums durch durch KI generierten Inhalt abzusichern.
- □ Überprüfen Sie bestehende KI-unterstützte Tools, die für die Code-Entwicklung verwendet werden. Die Fähigkeit eines Chatbots, Code zu schreiben, kann die Eigentumsrechte eines Unternehmens an seinem Produkt gefährden, wenn ein Chatbot zum Generieren von Code für das Produkt verwendet wird. Es könnte beispielsweise den Status und den Schutz des generierten Inhalts in Frage stellen und wer das Recht hat, den generierten Inhalt zu verwenden.
- □ Überprüfen Sie Risiken für das geistige Eigentum. Geistiges Eigentum, das von einem Chatbot generiert wird, könnte gefährdet sein, wenn bei der Generierung unrechtmäßig erlangte Daten verwendet wurden, die urheberrechtlich, markenrechtlich oder patentrechtlich geschützt sind. Wenn KI-Produkte rechtswidriges Material verwenden, besteht ein Risiko für die Ausgaben der KI, was zu einer Verletzung des geistigen Eigentums führen kann.
- □ Überprüfen Sie Verträge mit Haftungsbestimmungen. Haftungsausschlussklauseln versuchen, die Verantwortung für ein Ereignis, das zu Haftung führt, auf die Person zu legen, die dafür am schuldigsten war oder die die besten Chancen hatte, es zu stoppen. Legen Sie Leitplanken fest, um festzustellen, ob der Anbieter der KI oder sein Benutzer das Ereignis verursacht hat, das zu Haftung führt.
- □ Überprüfen Sie die Haftung für mögliche Verletzungen von Körper und Eigentum durch KI-Systeme.
- □ Überprüfen Sie die Versicherungsdeckung. Traditionelle (D&O) Haftpflicht- und allgemeine Haftpflichtversicherungen reichen wahrscheinlich nicht aus, um die Nutzung von KI vollständig abzudecken.
- □ Identifizieren Sie eventuelle Urheberrechtsprobleme. Urheberrecht erfordert menschliche Urheberschaft. Ein Unternehmen kann auch für Plagiate, Voreingenommenheit oder Verletzung des geistigen Eigentums haftbar gemacht werden, wenn LLM-Tools missbraucht werden.
- □ Stellen Sie sicher, dass Vereinbarungen für Auftragnehmer und die ordnungsgemäße Verwendung von KI für Entwicklung oder erbrachte Dienstleistungen vorhanden sind.
- □ Beschränken oder verbieten Sie die Verwendung von generativen KI-Tools für Mitarbeiter oder Auftragnehmer, bei denen durchsetzbare Rechte ein Problem darstellen können oder bei denen Bedenken hinsichtlich der Verletzung des geistigen Eigentums bestehen.
- □ Prüfen Sie AI-Lösungen, die für die Mitarbeiterverwaltung oder Einstellung verwendet werden. Diese könnten zu Klagen wegen ungleicher Behandlung oder ungleicher Auswirkungen führen.
- □ Stellen Sie sicher, dass die KI-Lösungen keine sensiblen Informationen ohne ordnungsgemäße Zustimmung oder Autorisierung erfassen oder weitergeben.

Regulatorische Aspekte

Der EU AI Act wird voraussichtlich das erste umfassende KI-Gesetz sein, wird jedoch frühestens 2025 in Kraft treten. Die Datenschutz-Grundverordnung (DSGVO) der EU geht zwar nicht speziell auf KI ein, enthält jedoch Regeln für die Datensammlung, Datensicherheit, Fairness und Transparenz, Genauigkeit und Zuverlässigkeit sowie Rechenschaftspflicht. Dies kann sich auf den Einsatz von GenAI auswirken. In den Vereinigten Staaten ist die KI-Regulierung in breitere Verbraucherschutzgesetze eingebunden. Zehn US-Bundesstaaten haben Gesetze verabschiedet oder werden bis Ende 2023 in Kraft treten.

Bundesbehörden wie die US-Gleichberechtigungskommission (EEOC), das Büro für Verbraucherschutz im Finanzwesen (CFPB), die Bundeshandelskommission (FTC) und die Abteilung für Bürgerrechte des US-Justizministeriums (DOJ) überwachen die Einhaltung von Fairness bei der Einstellung genau.

- □ Ermitteln Sie die spezifischen KI-Compliance-Anforderungen des Landes, des Bundesstaates oder anderen Regierungsbehörden.
- □ Ermitteln Sie die Compliance-Anforderungen für die Beschränkung der elektronischen Überwachung von Mitarbeitern und automatisierten Entscheidungssystemen im Zusammenhang mit der Beschäftigung
- □ Ermitteln Sie die Compliance-Anforderungen für die Zustimmung zur Gesichtserkennung und zur erforderlichen KI-Videoanalyse
- □ Überprüfen Sie KI-Tools, die bei der Einstellung oder Verwaltung von Mitarbeitern verwendet werden.
- □ Prüfen Sie die Einhaltung des Anbieters in Bezug auf geltende KI-Gesetze und bewährte Verfahren.
- □ Fragen Sie nach und dokumentieren Sie alle Produkte, die bei der Einstellung KI verwenden. Fragen Sie, wie das Modell trainiert wurde, wie es überwacht wird, und verfolgen Sie alle Korrekturen, um Diskriminierung und Voreingenommenheit zu vermeiden.
- □ Fragen Sie nach und dokumentieren Sie, welche Hostingoptionen enthalten sind.
- □ Fragen Sie nach und dokumentieren Sie, ob der Anbieter vertrauliche Daten sammelt.
- □ Fragen Sie, wie der Anbieter oder das Tool Daten speichert und löscht und den Einsatz von Gesichtserkennungs- und Videoanalysetools während der Vorbeschäftigung regelt.
- □ Überprüfen Sie andere organisationsbezogene gesetzliche Anforderungen an KI, die zu Compliance-Problemen führen können. Der Employee Retirement Income Security Act von 1974 beispielsweise hat Treuhandpflichtanforderungen für Rentenpläne, die ein Chatbot möglicherweise nicht erfüllen kann.

Verwendung oder Implementierung von Large Language Model-Lösungen

- □ Modellieren Sie die Bedrohungen und Architekturvertrauensgrenzen von LLM-Komponenten.
- □ Datensicherheit: Überprüfen Sie, wie Daten je nach Sensitivität klassifiziert und geschützt werden, einschließlich persönlicher und proprietärer Unternehmensdaten. (Wie werden Benutzerberechtigungen verwaltet und welche Sicherungsmaßnahmen sind vorhanden?)
- □ Zugangskontrolle: Implementieren Sie Zugriffskontrollen mit minimalen Berechtigungen und mehrere Sicherheitsebenen.
- □ Sicherheit der Trainingspipelines: Legen Sie strenge Kontrollen für das Datenmanagement, Pipelines, Modelle und Algorithmen während des Trainings fest.
- □ Sicherheit der Eingabe und Ausgabe: Überprüfen Sie Methoden zur Eingabevalidierung sowie zur Filterung, Bereinigung und Genehmigung von Ausgaben.
- □ Monitoring und Responses: Erstellen Sie Workflows, Überwachung und Reaktionen, um Automatisierung, Protokollierung und Prüfung zu verstehen. Bestätigen Sie, dass Auditprotokolle sicher sind.
- □ Führen Sie Anwendungstests, Quellcode-Überprüfungen, Schwachstellenanalysen und Red Teaming im Produktionsfreigabeprozess durch.
- □ Prüfen Sie auf vorhandene Sicherheitslücken im LLM-Modell oder in der Lieferkette.
- □ Untersuchen Sie die Auswirkungen von Bedrohungen und Angriffen auf LLM-Lösungen, wie z.B. Prompt-Injektion, Veröffentlichung sensibler Informationen und Prozessmanipulation.
- □ Untersuchen Sie die Auswirkungen von Angriffen und Bedrohungen auf LLM-Modelle, einschließlich Model-Poisoning, unsachgemäßer Datenverarbeitung, Angriffen auf die Lieferkette und Modell-Diebstahl.
- □ Sicherheit der Lieferkette: Fordern Sie Audits von Drittanbietern, Penetrationstests und Codeüberprüfungen für Drittanbieter an. (sowohl zu Beginn als auch kontinuierlich)
- □ Sicherheit der Infrastruktur: Fragen Sie, wie oft ein Anbieter Resilienztests durchführt? Was sind ihre SLAs in Bezug auf Verfügbarkeit, Skalierbarkeit und Leistung?
- □ Aktualisieren Sie die Incident-Response-Playbooks und nehmen Sie einen LLM-Zwischenfall in die Tischübung auf.
- □ Identifizieren oder erweitern Sie Metriken, um generative KI für Cybersicherheit mit anderen Ansätzen zu vergleichen, um erwartete Produktivitätsverbesserungen zu messen.

Test, Bewertung, Verifikation und Validierung (TEVV)

Der NIST AI Framework empfiehlt einen kontinuierlichen TEVV-Prozess während des gesamten KI-Lebenszyklus, an dem Betreiber des KI-Systems, Domänenexperten, KI-Designer, Benutzer, Produktentwickler, Evaluierer und Prüfer beteiligt sind. TEVV umfasst eine Reihe von Aufgaben wie Systemvalidierung, Integration, Tests, Neukalibrierung und kontinuierliche Überwachung für regelmäßige Aktualisierungen, um mit den Risiken und Veränderungen des KI-Systems umzugehen.

- □ Etablieren Sie kontinuierliche Tests, Bewertungen, Verifikationen und Validierungen während des gesamten KI-Modell-Lebenszyklus.
- □ Stellen Sie regelmäßige Executive-Metriken und Updates zur Funktionalität, Sicherheit, Zuverlässigkeit und Robustheit des KI-Modells bereit.

Modell- und Risikokarten

Modell- und Risikokarten sind grundlegende Elemente zur Steigerung der Transparenz, Rechenschaftspflicht und ethischen Bereitstellung von Large Language Models (LLMs). Modellkarten helfen Benutzern, KI-Systeme zu verstehen und ihnen zu vertrauen, indem sie standardisierte Dokumentationen über deren Design, Fähigkeiten und Einschränkungen bereitstellen, die zu informierten und sicheren Anwendungen führen. Risikokarten ergänzen dies, indem sie potenzielle negative Auswirkungen wie Vorurteile, Datenschutzprobleme und Sicherheitslücken offen ansprechen, was einen proaktiven Ansatz zur Vermeidung von Schäden fördert. Diese Dokumente sind gleichermaßen wichtig für Entwickler, Benutzer, Regulierungsbehörden und Ethiker, da sie eine kooperative Atmosphäre schaffen, in der die sozialen Auswirkungen von KI sorgfältig angegangen und behandelt werden. Diese von den Organisationen entwickelten und gepflegten Karten spielen eine wichtige Rolle dafür, dass KI-Technologien ethischen Standards und gesetzlichen Anforderungen entsprechen und eine verantwortungsbewusste Forschung und Bereitstellung im KI-Ökosystem ermöglichen.

Modellkarten enthalten wichtige Merkmale, die mit dem ML-Modell verbunden sind:

- **Modellinformationen**: Grundlegende Informationen über das Modell, wie Name, Version und Typ (Neuronales Netzwerk, Entscheidungsbaum usw.) sowie der beabsichtigte Anwendungsfall.

- **Modellarchitektur**: Enthält eine Beschreibung der Struktur des Modells, wie Anzahl und Typ der Schichten, Aktivierungsfunktionen und andere wichtige architektonische Entscheidungen.

- **Trainingsdaten und -methodik**: Informationen über die zum Training des Modells verwendeten Daten, wie Größe des Datensatzes, Datenquellen und verwendete Vorverarbeitungs- oder Datenaugmentationstechniken. Es enthält auch Details zur Trainingsmethodik, wie den verwendeten Optimierer, die Verlustfunktion und etwaige abgestimmte Hyperparameter.

- **Leistungsmetriken**: Informationen zur Leistung des Modells in Bezug auf verschiedene Metriken wie Genauigkeit, Präzision, Rückruf und F1-Score. Es kann auch Informationen darüber enthalten, wie das Modell auf verschiedenen Teilmengen der Daten abschneidet.

- **Potenzielle Vorurteile und Einschränkungen**: Listet potenzielle Vorurteile oder Einschränkungen des Modells auf, wie unbalancierte Trainingsdaten, Überanpassung oder Vorurteile in den Vorhersagen des Modells. Es kann auch Informationen über die Einschränkungen des Modells enthalten, wie seine Fähigkeit zur Verallgemeinerung auf neue Daten oder seine Eignung für bestimmte Anwendungsfälle.

- **Verantwortungsvolle KI-Überlegungen**: Alle ethischen oder verantwortungsvollen KI-Überlegungen im Zusammenhang mit dem Modell, wie Datenschutzbedenken, Fairness und Transparenz oder potenzielle gesellschaftliche Auswirkungen der Verwendung des Modells. Es kann auch Empfehlungen für weitere Tests, Validierung oder Überwachung des Modells enthalten.

Die in einer Modellkarte enthaltenen Merkmale können je nach Kontext und beabsichtigter Verwendung des Modells unterschiedlich sein, aber der Zweck besteht darin, Offenheit und Rechenschaftspflicht bei der Erstellung und Deployment von Machine-Learning-Modellen zu gewährleisten.

- □ Überprüfen Sie die Modellkarten der Modelle
- □ Überprüfen Sie die Risikokarte falls vorhanden
- □ Richten Sie einen Prozess ein, um Modellkarten für jedes bereitgestellte Modell nachzuverfolgen und zu pflegen, einschließlich der Modelle, die über einen Drittanbieter verwendet werden.

RAG: Optimierung großer Sprachmodelle

Fine-Tuning, die traditionelle Methode zur Optimierung eines vortrainierten Modells, besteht darin, ein vorhandenes Modell auf neuen, domänenspezifischen Daten neu zu trainieren und es für die Leistung in einer Aufgabe oder Anwendung anzupassen. Fine-Tuning ist teuer, aber notwendig, um die Leistung zu verbessern.

Retrieval-Augmented Generation (RAG) hat sich als effektivere Methode zur Optimierung und Erweiterung der Fähigkeiten großer Sprachmodelle entwickelt, indem relevante Daten aus aktuellen verfügbaren Wissensquellen abgerufen werden. RAG kann für bestimmte Domänen angepasst werden, um die Abfrage von domänenspezifischen Informationen zu optimieren und den Generierungsvorgang an die Nuancen spezialisierter Bereiche anzupassen. RAG gilt als effizientere und transparentere Methode zur Optimierung von LLMs, insbesondere für Probleme, bei denen gekennzeichnete Daten begrenzt oder teuer zu sammeln sind. Einer der Hauptvorteile von RAG ist die Unterstützung des kontinuierlichen Lernens, da neue Informationen kontinuierlich bei der Abrufphase aktualisiert werden können.

Die RAG-Implementierung umfasst mehrere Schlüsselschritte, angefangen von der Einbettung der Modellbereitstellung über die Indizierung der Wissensbibliothek bis hin zum Abrufen der relevantesten Dokumente für die Abfrageverarbeitung. Die effiziente Abfrage von relevantem Kontext basiert auf Vektor-Datenbanken, die für die Speicherung und Abfrage von Dokumenteneinbettungen verwendet werden.

RAG-Referenz

- □ Retrieval Augmented Generation (RAG) & LLM: Beispiele
- □ 12 RAG-Probleme und vorgeschlagene Lösungen

AI Red Teaming

AI Red Teaming ist eine adversarielle Angriffstestsimulation des KI-Systems, um zu überprüfen, ob vorhandene Schwachstellen von einem Angreifer ausgenutzt werden können. Es handelt sich um eine empfohlene Praxis vieler regulatorischer und KI-Governance-Stellen, einschließlich der Biden-Regierung. Red Teaming allein ist keine umfassende Lösung, um alle realen Schäden im Zusammenhang mit KI-Systemen zu validieren, und sollte zusammen mit anderen Formen von Tests, Bewertungen, Verifikationen und Validierungen wie algorithmischen Auswirkungsabschätzungen und externen Überprüfungen durchgeführt werden.

□ Integrieren Sie Red-Team-Tests als Standardverfahren für KI-Modelle und -Anwendungen.

Ressourcen

OWASP Top 10 für große Sprachmodellanwendungen

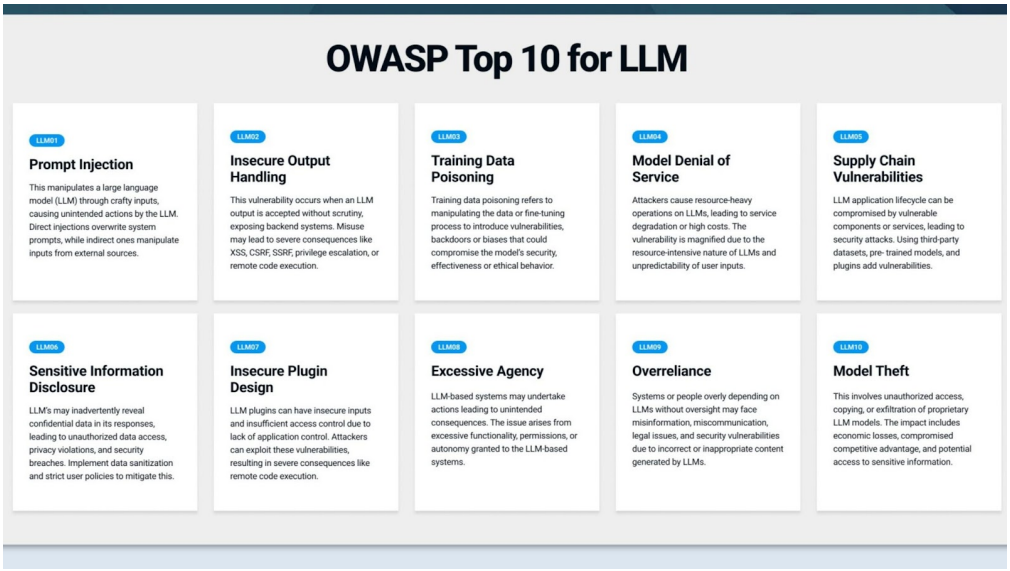

Figure 4.1: Bild der OWASP Top 10 für große Sprachmodellanwendungen

OWASP Top 10 für große Sprachmodellanwendungen visualisiert

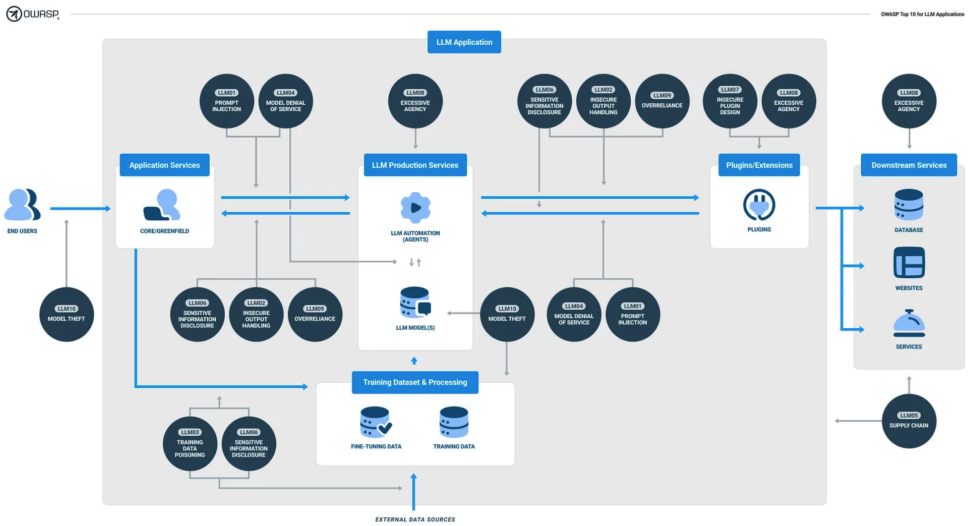

Figure 4.2: Bild der OWASP Top 10 für große Sprachmodellanwendungen visualisiert

OWASP-Ressourcen Die Verwendung von LLM-Lösungen erweitert die Angriffsfläche einer Organisation und bringt neue Herausforderungen mit sich, die spezielle Taktiken und Verteidigungsmaßnahmen erfordern. Es stellt auch Probleme dar, die ähnlich wie bekannte Probleme sind, und es gibt bereits etablierte IT-Sicherheitsverfahren und -minderungsmaßnahmen. Die Integration der LLM-IT-Sicherheit mit den etablierten IT-Sicherheitskontrollen, -prozessen und -verfahren einer Organisation ermöglicht es einer Organisation, ihre Anfälligkeit gegenüber Bedrohungen zu verringern. Wie sie sich gegenseitig integrieren, ist unter den OWASP-Integrationsstandards verfügbar.

OWASP-Ressource	Beschreibung	Warum es empfohlen wird und wo es verwendet werden soll
OWASP SAMM	Software Assurance Maturity Model (Modell zur Reifegradbestimmung der Software-Sicherheit)	Bietet eine effektive und messbare Möglichkeit, den sicheren Entwicklungslebenszyklus einer Organisation zu analysieren und zu verbessern. SAMM unterstützt den gesamten Software-Lebenszyklus. Es ist iterativ und risikoorientiert und ermöglicht es Organisationen, Lücken in der sicheren Softwareentwicklung zu identifizieren und zu priorisieren, damit Ressourcen zur Verbesserung des Prozesses dort eingesetzt werden können, wo sie den größten Verbesserungseffekt haben.
OWASP AI Security and Privacy Guide	OWASP-Projekt mit dem Ziel, weltweit den Austausch über KI-Sicherheit zu fördern, die Ausrichtung von Standards zu unterstützen und die Zusammenarbeit voranzutreiben.	Der OWASP AI Security and Privacy Guide ist eine umfassende Liste der wichtigsten KI-Sicherheits- und Datenschutzaspekte. Er soll eine umfassende Ressource für Entwickler, Sicherheitsforscher und Sicherheitsberater sein, um die Sicherheit und den Datenschutz von KI-Systemen zu überprüfen.
OWASP AI Exchange	OWASP AI Exchange ist die Methode zur Erfassung von Informationen für den OWASP AI Security and Privacy Guide.	Der AI Exchange ist die primäre Methode zur Erfassung von Informationen, die von OWASP verwendet wird, um die Ausrichtung des OWASP AI Security and Privacy Guide zu steuern.

OWASP-Ressource	Beschreibung	Warum es empfohlen wird und wo es verwendet werden soll
OWASP Machine Learning Security Top 10	OWASP Machine Learning Security Top 10-Sicherheitsprobleme von KI-Systemen.	Das OWASP Machine Learning Security Top 10 ist eine gemeinschaftliche Anstrengung, um die wichtigsten Sicherheitsprobleme von KI-Systemen in einer Form zu sammeln und darzustellen, die sowohl von Sicherheitsexperten als auch von Datenwissenschaftlern leicht verständlich ist. Dieses Projekt enthält die ML Top 10 und ist ein laufendes Arbeitsdokument, das klare und handlungsorientierte Einblicke in das Design, die Erstellung, das Testen und die Beschaffung von sicheren und datenschutzorientierten KI-Systemen bietet. Es ist die beste OWASP-Ressource für Informationen zu globalen KI-Vorschriften und zum Datenschutz.
OpenCRE	OpenCRE (Common Requirement Enumeration) ist die interaktive Content-Linking-Plattform zur Zusammenführung von Sicherheitsstandards und -richtlinien in einer Übersicht.	Verwenden Sie diese Website, um nach Standards zu suchen. Sie können nach Standardnamen oder nach Steuerungstyp suchen.
OWASP Threat Modeling	Ein strukturierter, formaler Prozess für das Bedrohungsmodellieren einer Anwendung	Erfahren Sie alles über das Bedrohungsmodellieren, das eine strukturierte Darstellung aller Informationen ist, die die Sicherheit einer Anwendung beeinflussen.

OWASP-Ressource	Beschreibung	Warum es empfohlen wird und wo es verwendet werden soll
OWASP CycloneDX	OWASP CycloneDX ist ein umfassender Standard für die vollständige Auflistung aller Komponenten einer Software (Bill of Materials, BOM) und bietet erweiterte Funktionen zur Reduzierung von Cyberrisiken in der Lieferkette.	Moderne Software wird unter Verwendung von Komponenten von Drittanbietern und Open Source zusammengesetzt. Sie werden auf komplexe und einzigartige Weise miteinander verbunden und mit eigenem Code integriert, um die gewünschte Funktionalität zu erreichen. Ein SBOM stellt ein genaues Inventar aller Komponenten bereit, mit dem Organisationen Risiken identifizieren können. Es ermöglicht eine größere Transparenz und ermöglicht eine schnelle Auswirkungsanalyse. EO 14028 legte Mindestanforderungen für SBOM für Bundesbehörden fest.
OWASP Software Component Verification Standard (SCVS)	Eine gemeinschaftliche Anstrengung zur Entwicklung eines Frameworks zur Identifizierung von Aktivitäten, Kontrollen und bewährten Verfahren, die zur Identifizierung und Reduzierung von Risiken in einer Software-Lieferkette beitragen können.	Verwenden Sie SCVS, um einen gemeinsamen Satz von Aktivitäten, Kontrollen und bewährten Verfahren zu entwickeln, die Risiken in einer Software-Lieferkette reduzieren können, und um eine Grundlage und einen Weg zu einer ausgereiften Wachsamkeit in der Software-Lieferkette zu identifizieren.
OWASP API Security Project	API-Sicherheit konzentriert sich auf Strategien und Lösungen zur Identifizierung und Minderung der einzigartigen Schwachstellen und Sicherheitsrisiken von APIs (Application Programming Interfaces)	APIs sind ein grundlegendes Element für die Verbindung von Anwendungen, und die Minderung von Konfigurationsfehlern oder Schwachstellen ist erforderlich, um Benutzer und Organisationen zu schützen. Verwenden Sie es für die Sicherheitstests und das Red Teaming der Entwicklungs- und Produktionsumgebung.

OWASP-Ressource	Beschreibung	Warum es empfohlen wird und wo es verwendet werden soll
OWASP Application Security Verification Standard ASVS	Der Application Security Verification Standard (ASVS) bietet eine Grundlage für die Prüfung technischer Sicherheitskontrollen von Webanwendungen und bietet Entwicklern auch eine Liste von Anforderungen für die sichere Entwicklung.	Ein Cookbook für Sicherheitsanforderungen an Webanwendungen, Sicherheitstests und Metriken. Verwenden Sie es, um Sicherheitsbenutzerstorys und Sicherheitsanwendungsfälle für die Testfreigabe festzulegen.
OWASP Threat and Safeguard Matrix (TaSM)	Eine handlungsorientierte Sichtweise zum Schutz und zur Förderung des Unternehmens	Diese Matrix ermöglicht es einem Unternehmen, seine wichtigsten Bedrohungen mit den NIST Cyber Security Framework Functions (Identifizieren, Schützen, Erkennen, Reagieren und Wiederherstellen) zu verbinden, um einen robusten Sicherheitsplan zu erstellen. Verwenden Sie sie als Dashboard, um Sicherheit in der gesamten Organisation zu verfolgen und zu berichten.
Defect Dojo	Ein Open-Source-Schwachstellenmanagement-Tool, das den Testprozess durch Vorlagen, Berichterstellung, Metriken und Self-Service-Tools optimiert.	Verwenden Sie Defect Dojo, um die Zeit für die Erfassung von Schwachstellen zu reduzieren, indem Sie Vorlagen für Schwachstellen, Importe für gängige Schwachstellenscanner, Berichterstellung und Metriken bereitstellen.

Table 4.1: OWASP-Ressourcen

MITRE Ressourcen Die erhöhte Häufigkeit von LLM-Bedrohungen unterstreicht den Wert eines resilienzorientierten Ansatzes zur Verteidigung der Angriffsfläche einer Organisation. Bestehende TTPs werden mit neuen Angriffsflächen und Fähigkeiten in LLM-Bedrohungen und -Abwehrmaßnahmen kombiniert. MITRE unterhält einen etablierten und weit verbreiteten Mechanismus zur Koordination gegnerischer Taktiken und Verfahren, basierend auf realen Beobachtungen.

Die Koordination und Zuordnung der LLM-Sicherheitsstrategie einer Organisation zu MITRE ATT&CK und MITRE ATLAS ermöglicht es einer Organisation festzustellen, wo die LLM-Sicherheit durch bestehende Prozesse wie API-Sicherheitsstandards abgedeckt ist oder wo Sicherheitslücken bestehen.

MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) ist ein Framework, eine Sammlung von Datenmatrizen und ein Bewertungswerkzeug, das von der MITRE Corporation entwickelt wurde, um Organisationen dabei zu helfen, die Wirksamkeit ihrer Cybersicherheit über ihre gesamte digitale Angriffsfläche hinweg zu bewerten und bisher unentdeckte Lücken zu finden. Es handelt sich um ein weltweit genutztes Wissensrepository. Die MITRE ATT&CK-Matrix enthält eine Sammlung von Strategien, die von Angreifern verwendet werden, um ein bestimmtes Ziel zu erreichen. In der ATT&CK-Matrix werden diese Ziele als Taktiken klassifiziert. Die Ziele werden in Angriffsreihenfolge dargestellt, beginnend mit der Aufklärung und endend mit der letztendlichen Zielsetzung der Exfiltration oder Auswirkung.

MITRE ATLAS, was für "Adversarial Threat Landscape for Artificial Intelligence Systems" steht, ist eine Wissensbasis, die auf realen Beispielen von Angriffen auf maschinelles Lernen (ML) durch böswillige Akteure basiert. ATLAS basiert auf der MITRE ATT&CK-Architektur, und seine Taktiken und Verfahren ergänzen diejenigen, die in ATT&CK zu finden sind.

MITRE Ressource	Beschreibung	Warum wird es empfohlen und wo wird es verwendet
MITRE ATT&CK	Wissensbasis von gegnerischen Taktiken und Techniken basierend auf realen Beobachtungen	Die ATT&CK-Wissensbasis wird als Grundlage für die Entwicklung spezifischer Bedrohungsmodelle und -methoden verwendet. Ordnen Sie bestehende Kontrollen in der Organisation gegnerischen Taktiken und Techniken zu, um Lücken oder Bereiche für Tests zu identifizieren.
MITRE AT&CK Workbench	Erstellen oder erweitern Sie ATT&CK-Daten in einer lokalen Wissensbasis	Hosten und verwalten Sie eine angepasste Kopie der ATT&CK-Wissensbasis. Diese lokale Kopie der ATT&CK-Wissensbasis kann um neue oder aktualisierte Techniken, Taktiken, Gruppen von Abwehrmaßnahmen und Software erweitert werden, die spezifisch für Ihre Organisation sind.

MITRE Ressource	Beschreibung	Warum wird es empfohlen und wo wird es verwendet
MITRE ATLAS	MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) ist eine Wissensbasis von gegnerischen Taktiken, Techniken und Fallstudien für maschinelles Lernen (ML) basierend auf realen Beobachtungen, Demonstrationen von ML-Red Teams und Sicherheitsgruppen sowie dem Stand des Möglichen aus akademischer Forschung	Verwenden Sie es, um bekannte ML-Schwachstellen abzugleichen und Checks und Kontrollen für vorgeschlagene Projekte oder bestehende Systeme abzubilden.
MITRE ATT&CK Powered Suit	ATT&CK Powered Suit ist eine Browser-Erweiterung, mit der die MITRE ATT&CK-Wissensbasis jederzeit verfügbar ist.	Fügen Sie es Ihrem Browser hinzu, um schnell nach Taktiken, Techniken und mehr zu suchen, ohne Ihren Arbeitsablauf zu unterbrechen.
The Threat Report ATT&CK Mapper (TRAM)	Automatisiert die Identifizierung von TTPs in CTI-Berichten	Die Zuordnung von TTPs in CTI-Berichten zu MITRE ATT&CK ist schwierig, fehleranfällig und zeitaufwändig. TRAM verwendet LLMs, um diesen Prozess für die 50 häufigsten Techniken zu automatisieren. Unterstützt Jupyter-Notebooks.
Attack Flow v2.1.0	Attack Flow ist eine Sprache zur Beschreibung, wie Cyber-Angreifer verschiedene offensive Techniken kombinieren und sequenzieren, um ihre Ziele zu erreichen.	Attack Flow hilft dabei zu visualisieren, wie ein Angreifer eine Technik verwendet, sodass Verteidiger und Führungskräfte verstehen, wie Angreifer vorgehen, und ihre eigene Verteidigungsstrategie verbessern können.

MITRE Ressource	Beschreibung	Warum wird es empfohlen und wo wird es verwendet
MITRE Caldera	Eine Cybersicherheitsplattform (Framework), die die automatisierte Emulation von Angreifern ermöglicht, manuelle Red Teams unterstützt und die automatisierte Reaktion auf Vorfälle ermöglicht.	Plugins für Caldera stehen zur Verfügung, um die Kernfunktionen des Frameworks zu erweitern und zusätzliche Funktionen bereitzustellen, einschließlich Agenten, Berichterstattung, Sammlungen von TTPs und anderem.
CALDERA Plugin: Arsenal	Ein Plugin, das für die Emulation von Angriffen auf KI-gestützte Systeme entwickelt wurde.	Dieses Plugin stellt in MITRE ATLAS definierte TTPs zur Verfügung, um mit CALDERA zu interagieren.
Atomic Red Team	Eine Bibliothek von Tests, die dem MITRE ATT&CK-Framework zugeordnet sind.	Verwenden Sie es, um Kontrollen in einer Umgebung zu validieren und zu testen. Sicherheitsteams können Atomic Red Team verwenden, um ihre Umgebungen schnell, portabel und reproduzierbar zu testen. Sie können atomare Tests direkt von der Kommandozeile ausführen, es ist keine Installation erforderlich.
MITRE CTI Blueprints	Automatisiert die Erstellung von Cyber Threat Intelligence-Berichten.	CTI Blueprints helfen Cyber Threat Intelligence (CTI)-Analysten dabei, hochwertige, handlungsorientierte Berichte konsistenter und effizienter zu erstellen.

Table 4.2: MITRE Ressourcen

AI Schwachstellen-Repositorys

Name (Name)	Beschreibung
AI Incident Database (KI-Zwischenfall-Datenbank)	Ein Repository von Artikeln über verschiedene Fälle, in denen KI in realen Anwendungen versagt hat, gepflegt von einer Forschungsgruppe eines Colleges und durch Crowdsourcing.
OECD AI Incidents Monitor (AIM) (OECD KI-Zwischenfall-Überwachung)	Bietet einen zugänglichen Ausgangspunkt, um die Landschaft der mit KI verbundenen Herausforderungen zu verstehen.
Drei der führenden Unternehmen, die KI-Modell-Schwachstellen verfolgen	
Huntr Bug Bounty : ProtectAI	Bug-Bounty-Plattform für KI/ML
AI Vulnerability Database (AVID) : Garak	Datenbank für Modellschwachstellen
AI Risk Database: Robust Intelligence	Datenbank für Modellschwachstellen

Table 4.3: KI-Schwachstellen-Repositorys

Künstliche Intelligenz (KI) Beschaffungsrichtlinien

Name	Beschreibung
Weltwirtschaftsforum: Verantwortungsvolle Anwendung von KI: Richtlinien für die Beschaffung von KI-Lösungen durch den privaten Sektor: Insight-Bericht Juni 2023	Die Standard-Benchmarks und Bewertungskriterien für den Erwerb von Künstliche Intelligenz (KI)-Systemen befinden sich noch in der frühen Entwicklung. Die Beschaffungsrichtlinien bieten Organisationen eine Grundlage für die Berücksichtigung des gesamten Beschaffungsprozesses. Verwenden Sie diese Richtlinien, um den bestehenden Prozess der Beschaffung von Lieferanten und Anbietern im Bereich Drittrisiko zu ergänzen.

Table 4.4: KI Beschaffungsrichtlinien

Team

Dank and die Mitwirkenden der OWASP Top 10 for LLM Applications Cybersecurity und Governance Checklist.

Mitwirkende der Checkliste		
Sandy Dunn	Heather Linn	John Sotiropoulos
Steve Wilson	Fabrizio Cilli	Aubrey King
Bob Simonoff	David Rowe	Rob Vanderveer
Emmanual Guilherme Junior	Andrea Succi	Jason Ross
Talesh Seeparsan	Anthony Glynn	Julie Tao

Table A.1: OWASP LLM AI Security & Governance Checklist Team