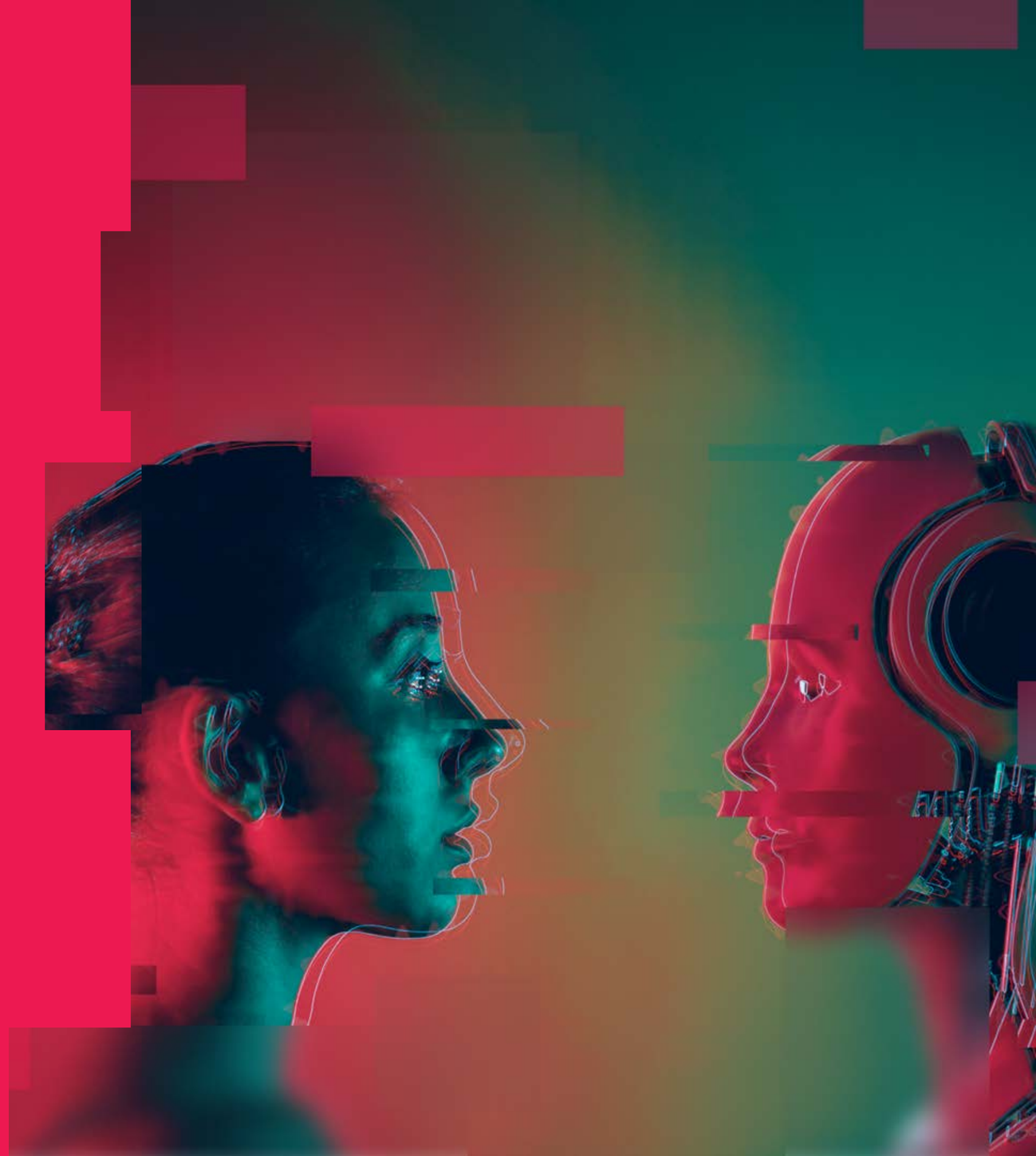
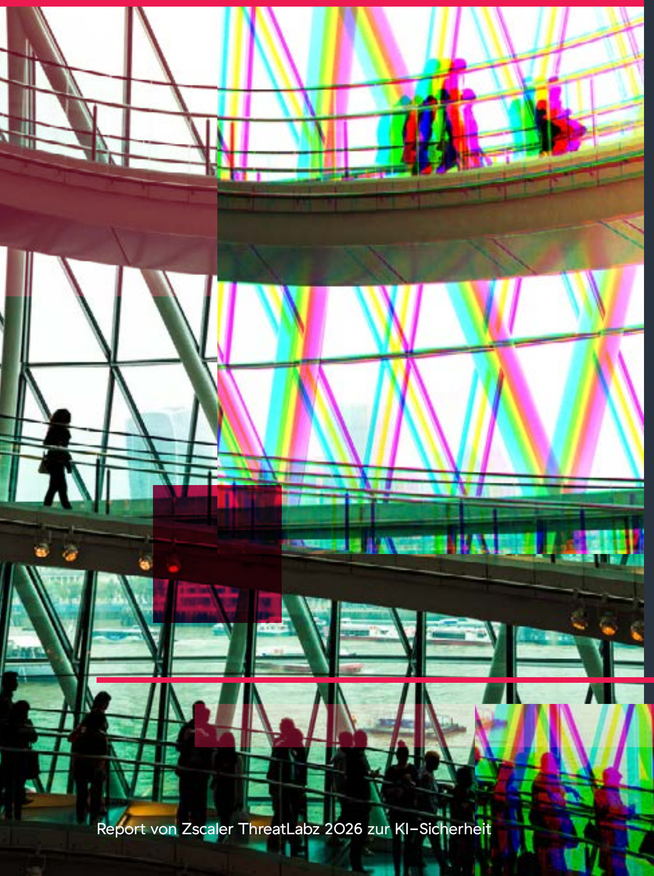
ThreatLabz-Report 2026
zur KI-Sicherheit



Inhaltsverzeichnis



Kurzfassung_

Wer 2025 mit KI arbeitete, erlebte eine Welt, die von enormem Tempo, einer riesigen Flut an neuen Tools und unaufhörlicher Dynamik geprägt war.

Unternehmen setzen KI und maschinelles Lernen (KI/ML) in allen Geschäftsbereichen ein, um schneller zu handeln, Entscheidungen zu automatisieren und die Produktivität zu steigern. KI unterstützt Entwicklung, Kommunikation, Forschung und operative Abläufe in einem Tempo, das vor wenigen Jahren noch undenkbar gewesen wäre. Diese Dynamik bringt jedoch neue Herausforderungen mit sich: Immer mehr sensible Daten fließen durch eine wachsende Zahl von KI-/ML-Anwendungen — häufig mit eingeschränkter Transparenz und weniger Schutzmechanismen.

Der zunehmende Einsatz von KI erweitert die Angriffsfläche von Unternehmen, und Angreifer haben in den letzten zwölf Monaten schnell reagiert. Niedrigere Barrieren und realistischere Szenarien machen Angriffe schneller und glaubwürdiger, während frühe Anzeichen für den Missbrauch agentischer und halbautonomer KI auf eine Verschiebung in der Bedrohungslandschaft hinwiesen. Gleichzeitig sehen sich Organisationen mit einer wachsenden Vielfalt an Risiken konfrontiert — von Schatten- und eingebetteter KI über Halluzinationen bis hin zu ungeschützten privaten Modellen.

Wie können Unternehmen Umgebungen absichern, in denen KI allgegenwärtig ist, KI-gestütze Innovationen fördern und sich gleichzeitig gegen KI-basierte Bedrohungen zur Wehr setzen — ohne das Geschäft auszubremsen?

Der Report von Zscaler ThreatLabz zur KI-Sicherheit 2026 beleuchtet, wie Unternehmen genau diesen Balanceakt meistern. Der Report basiert auf der Analyse von 989,3 Milliarden KI-/ML-Transaktionen, die zwischen Januar 2025 und Dezember 2025 über die Zscaler Zero Trust Exchange™ beobachtet wurden, und liefert damit einen fundierten Einblick, wie KI weltweit tatsächlich eingesetzt und kontrolliert wird.

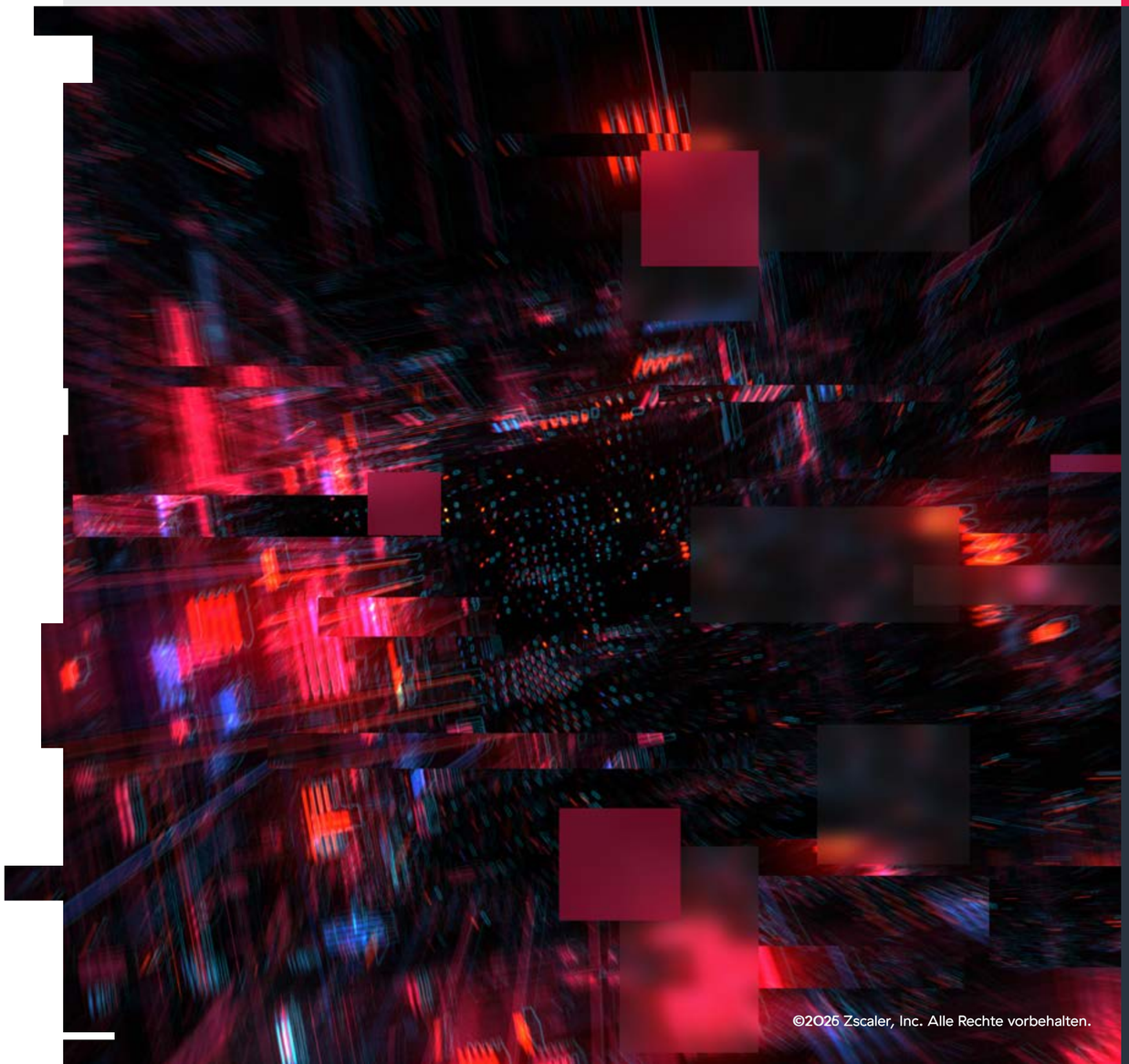
Die Daten zeigen eine anhaltende Beschleunigung: Die Aktivität von Unternehmen im Bereich KI/ML stieg im Jahresvergleich um 83.3 %, während das Datenvolumen um 92,6 % auf über 18.000 Terabyte (TB) anwuchs. Auf dieser Ebene funktioniert KI weniger als eine Reihe einzelner Tools, sondern eher wie eine dauerhaft aktive Infrastruktur, die Unternehmensdaten kontinuierlich bewegt und verarbeitet. Der Zugriff ist jedoch nach wie vor eingeschränkt: Unternehmen blockierten 39 % der KI-/ML-Transaktionen, was die anhaltenden Bedenken hinsichtlich Datenexposition, Datenschutz und Richtliniendurchsetzung widerspiegelt.

Die Nutzungsmuster zeigen auch, wo Nutzen und Risiko aufeinandertreffen. Die KI-Anwendungen, auf die Mitarbeitende am meisten angewiesen sind — etwa Codeium, Grammarly und ChatGPT — stehen im Zentrum der Arbeitsprozesse. Sie erzeugen die höchsten Aktivitätslevel und markieren zugleich die kritischen Punkte in unseren Risikoanalysen.

Im Jahr 2026 geht es beim Schutz von KI um mehr als die Kontrolle einzelner KI-/ML-Anwendungen. Entscheidend ist, wie KI im gesamten Unternehmen erfasst, entwickelt, genutzt und gesteuert wird. Organisationen benötigen Transparenz über KI-Nutzung und Risiken, Schutzmechanismen, die KI-Systeme und -Daten in Echtzeit schützen, sowie konsequente Kontrollen, die den Zugriff absichern und gleichzeitig Innovation ermöglichen. Dieser Report liefert Einblicke in die Trends, Herausforderungen und Realitäten, die KI-Sicherheit prägen, und Handlungsempfehlungen für eine sichere Einführung von KI.

Was Führungskräfte im Unternehmen jetzt wissen müssen

- **KI ist inzwischen Teil der Unternehmensinfrastruktur.** Fast eine Billion Transaktionen verdeutlichen die permanente, durchgängige Nutzung dieser Systeme. KI muss mit derselben Sorgfalt wie Cloud, Identitäten und Daten gesteuert werden, um eine sichere und skalierbare Nutzung zu gewährleisten.
- **Das Risiko von Datenlecks hängt heute weniger von bösen Absichten ab als vielmehr von der schieren Masse.** Wenn Daten im Petabyte-Bereich durch KI-Workflows fließen, steigt die Exposition durch hohe Geschwindigkeit und Wiederholungen, selbst bei genehmigter Nutzung im Einklang mit den Unternehmenszielen.
- **Genehmigte KI-Tools bilden die primäre Angriffsfläche.** Gängige, genehmigte KI-Tools machen den Großteil der KI-Aktivitäten und Dateninteraktionen in Unternehmen aus. Schatten-KI bleibt zwar ein zentrales Problem, doch allein gegen nicht autorisierte Tools vorzugehen reicht nicht aus, um alle KI-bezogenen Risiken und Gefahren zu mindern.
- **Sicherheitsvorgaben schränken die KI-Einführung ein.** Da 39 % der KI-Transaktionen blockiert werden, trägt die Durchsetzung von Richtlinien aktiv dazu bei, wie KI eingesetzt wird. Das zeigt, dass Governance funktioniert und dass Unternehmen ein Gleichgewicht zwischen Innovation und Sicherheit suchen.
- **Herkömmliche Sicherheitsmodelle passen nicht zu KI-Workflows.** Kontrollen, die für langsame menschliche Abläufe und statische Daten gedacht sind, kommen bei den rasanten, automatisierten KI-Aktivitäten nicht hinterher.
- **Wer KI im großen Maßstab steuern kann, verschafft sich einen Wettbewerbsvorteil.** Unternehmen, die den breiten Einsatz von KI mit starken, integrierten Kontrollen ermöglichen, werden schneller voran-kommen als diejenigen, die die Nutzung wegen unkontrollierter Risiken komplett einschränken müssen.



Haupter-kenntnisse

Im Zeitraum von Januar bis Dezember 2025 hat ThreatLabz **989,3 Milliarden KI- und ML-Transaktionen** in der Zscaler-Cloud ausgewertet. Die folgenden zentralen Erkenntnisse stützen sich auf Daten aus verschiedenen Zeiträumen*, die für eine vergleichende Analyse herangezogen wurden.

KI in Unternehmen befindet sich weiter auf steilem Wachstumskurs. Der Einsatz von KI- und ML-Anwendungen erhöhte sich gegenüber dem Vorjahr um 83 % und erreichte fast eine Billion Transaktionen in einem Ökosystem mit mehr als 3.400 Anwendungen.

Unternehmen senden immer größere Datenmengen an KI-Tools. Insgesamt übertrugen sie 18.033 TB an KI- und ML-Anwendungen, ein Anstieg um 93 % gegenüber dem Vorjahr.

Die große Zahl blockierter KI- und ML-Transaktionen deutet auf ein aktives Risikomanagement hin. Unternehmen blockierten 39 % aller KI- und ML-Transaktionen und zeigen damit, dass Sorgen rund um Datenexposition, Datenschutz und Richtlinienkonformität mit dem wachsenden KI-Einsatz nicht nachlassen.

KI in Unternehmen weist erhebliche Sicherheitslücken auf. Die Red-Teaming-Experten von Zscaler konnten die meisten Systeme bereits nach 16 Minuten kompromittieren und fanden in jedem einzelnen der getesteten Systeme schwerwiegende Schwachstellen.

* Datenerhebungszeiträume:

- Jahresanalyse und Vergleich zum Vorjahr: Januar bis Dezember 2025, inklusive Vergleich mit dem gleichen Zeitraum 2024.
- DLP-Verstöße und länderspezifische Daten: Juni bis Dezember 2025.

OpenAI ist führender Anbieter von LLM-Lösungen.
Der Großteil aller LLM-basierten Transaktionen in Unternehmen entfiel auf OpenAI — dreimal so viele wie auf Codeium — und macht OpenAI damit faktisch zum derzeitigen Standard.

ChatGPT verursacht den überwiegenden Teil aller DLP-Verstöße. In den untersuchten KI- und ML-Anwendungen erzeugte ChatGPT 410 Millionen Verstöße gegen Richtlinien zum Schutz vor Datenverlusten, was die Risiken unterstreicht, die Unternehmen bei der Nutzung kontextabhängiger KI-Assistenten tragen.

Integrierte Produktivitätsanwendungen bilden das Rückgrat für den Einsatz von KI in Unternehmen.
Grammarly wurde zur Nummer eins nach Transaktionsvolumen und zeigt, wie sehr Unternehmen auf KI setzen, die nahtlos in ihre Kommunikations- und Arbeitsabläufe eingebunden ist.

Finanz- und Versicherungswesen sowie die Fertigungsindustrie führen erneut beim Einsatz von KI. Zum dritten Jahr in Folge entfiel der größte Teil des KI- und ML-Traffics auf diese Sektoren (23 % und 20 %), was auf Modernisierungsmaßnahmen und aufwändige Dokumentationsabläufe zurückzuführen ist.

Die meisten KI- und ML-Transaktionen stammen weiterhin aus den USA. Mit 38 % des Gesamtvolumens führten die USA die Rangliste an, gefolgt von Indien (14 %) und Kanada (5 %).

Mit der wachsenden Nutzung von KI vergrößert sich die Angriffsfläche in Unternehmen. Je stärker KI in Geschäftsprozesse integriert wird, desto mehr Möglichkeiten entstehen, dass Daten oder Zugriffe offengelegt werden. Das erhöht das Risiko von Datenlecks, missbräuchlichem Einsatz von Prompts und KI-gestützten Angriffen — und macht Zero-Trust-Architekturen sowie KI-basierte Sicherheitskontrollen unverzichtbar.

Trends bei der KI-/ML-Nutzung

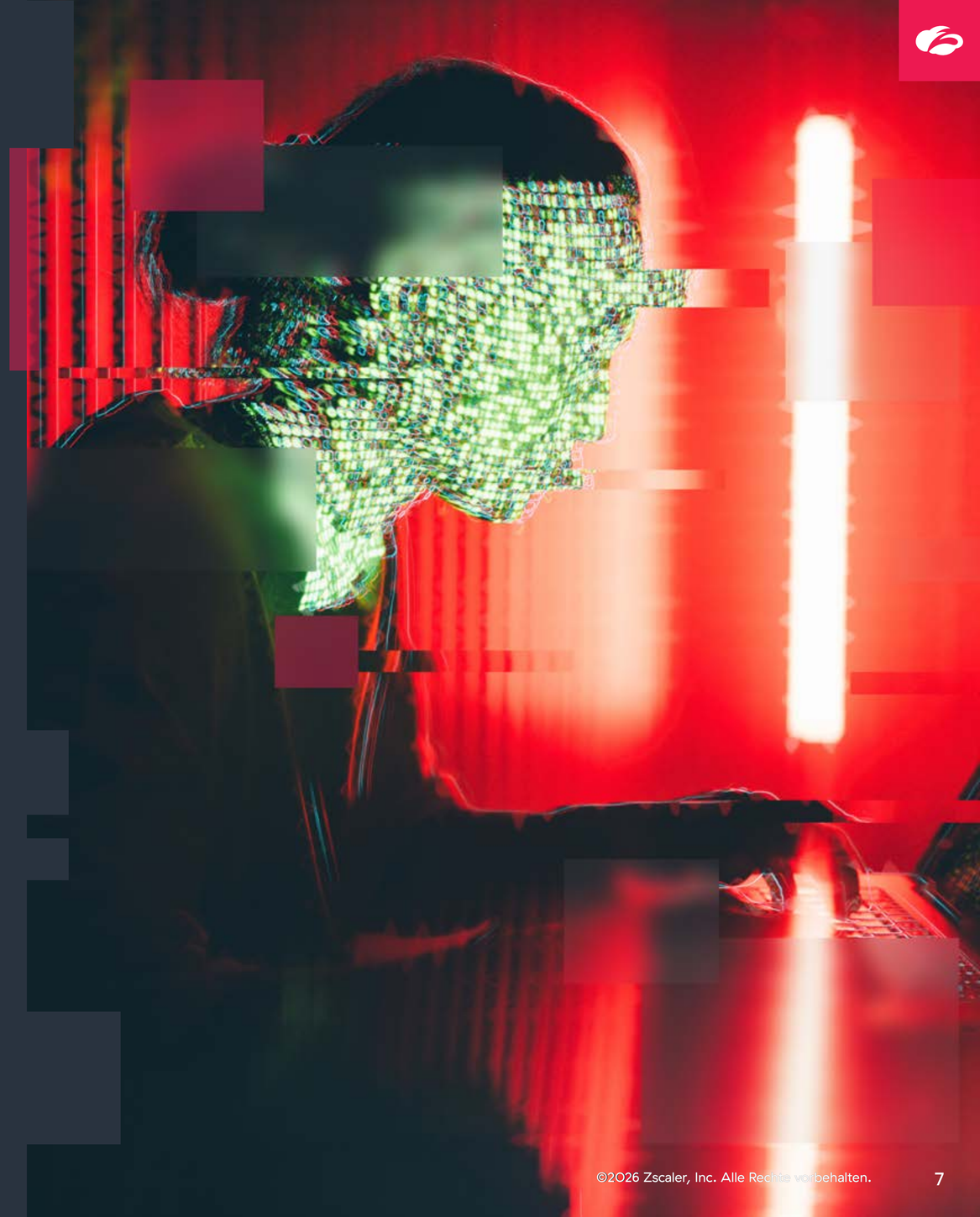
Der Einsatz von KI in Unternehmen setzte 2025 seinen steilen Aufwärtstrend fort.

ThreatLabz hat die Analyse der KI-Nutzung auf mehr als 3.400 Anwendungen ausgeweitet — viermal so viele wie noch im Vorjahr —, die KI- und ML-Transaktionen generieren. Viele dieser Apps erzeugen nur wenig Traffic, doch das schnelle Wachstum des gesamten Ökosystems ist ein klarer Hinweis darauf, wie rasant sich KI-Funktionen über Anbieter, Anwendungsfälle und Unternehmensbereiche hinweg verbreiten und damit sowohl neue Chancen als auch Risiken entstehen.

Um nachzuvollziehen, wie dieses Wachstum sich in der Praxis in Unternehmen auswirkt, analysierte ThreatLabz die KI- und ML-Aktivität auf mehreren Ebenen:

- **Alle KI/ML-Transaktionen:** Analyse nach URL-Kategorie, inklusive zugelassener und blockierter Aktivitäten.
- **LLM-Anbieter:** Ranking nach erzeugtem KI-/ML-Traffic und Einfluss auf Unternehmens-Workflows.
- **Führende KI-/ML-Anwendungen:** Identifikation der Apps mit dem größten Beitrag zur KI-Aktivität in Unternehmen.
- **KI-Nutzung in Abteilungen:** Zuordnung häufig genutzter Apps zu typischen Unternehmens-abteilungen, um Anwendungsbereiche im Tagesgeschäft zu erkennen.

Auf Basis dieser Analysen wollen wir aufzeigen, wie KI in Unternehmen wirklich genutzt wird und an welchen Stellen sich Nutzung, Abhängigkeit und Risiken überschneiden.



Globales Wachstum von KI-/ML-Transaktionen

Die Zahl der KI-/ML-Transaktionen näherte sich 2025 der Billionenmarke und erreichte insgesamt 989,3 Milliarden. Ein Großteil dieses Wachstums ist auf häufig genutzte Anwendungen wie ChatGPT, Grammarly und Codeium zurückzuführen.

TRENDS BEI DER KI-/ML-NUTZUNG NACH TRANSAKTIONSVOLUMEN

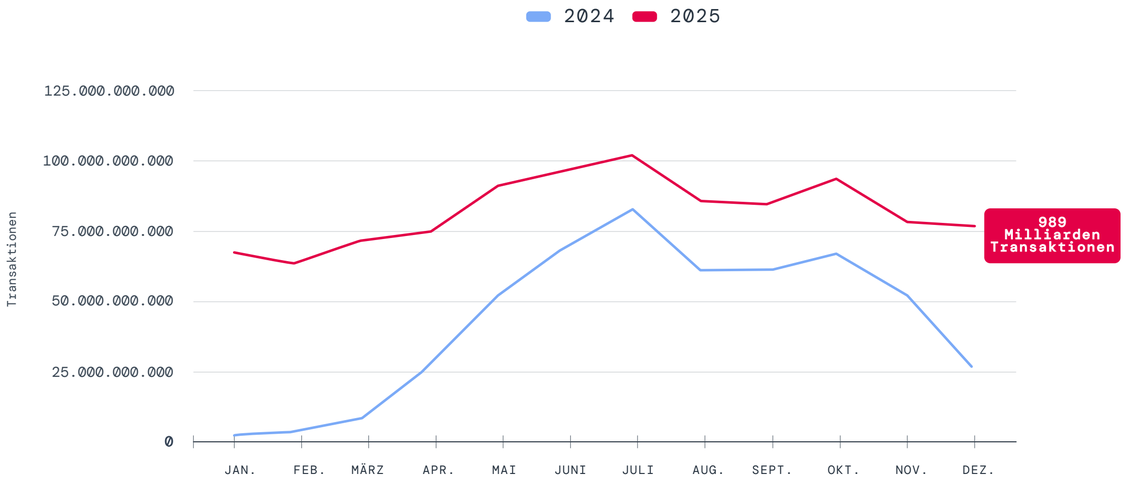


Abb. 1: KI-/ML-Transaktionen im Jahresvergleich (Januar-Dezember 2025)

WICHTIGES ERGEBNIS

Die KI-/ML-Aktivität stieg im Jahresvergleich um 83% und umfasste ein Ökosystem von mehr als 3.400 Anwendungen.

Wie schon in den Vorjahren fällt ein Teil des Traffics unter die Kategorie „Allgemeine KI-Anwendungen“. Dabei handelt es sich um KI-/ML-Transaktionen, die keiner bekannten Anwendung zugeordnet werden können, aber von Zscalers KI-/ML-gestützter URL-Kategorisierung als KI-relevant erkannt werden. Diese Klassifizierung analysiert Texte, Bilder und weitere Inhaltsmerkmale, um KI-Aktivitäten zu identifizieren. Da neue KI-Anwendungen schneller entstehen, als sie manuell erfasst werden können, ist es besonders wichtig, bisher unbekannte KI-Traffic-Quellen zu erkennen und in die Sicherheitsrichtlinien einzubinden.

Soweit nicht anders angegeben, fokussiert sich die weitere Analyse in diesem Report auf klassifizierte Anwendungen. So lässt sich nachvollziehen, wie KI in Unternehmen über etablierte KI-/ML-Anwendungen tatsächlich genutzt wird.

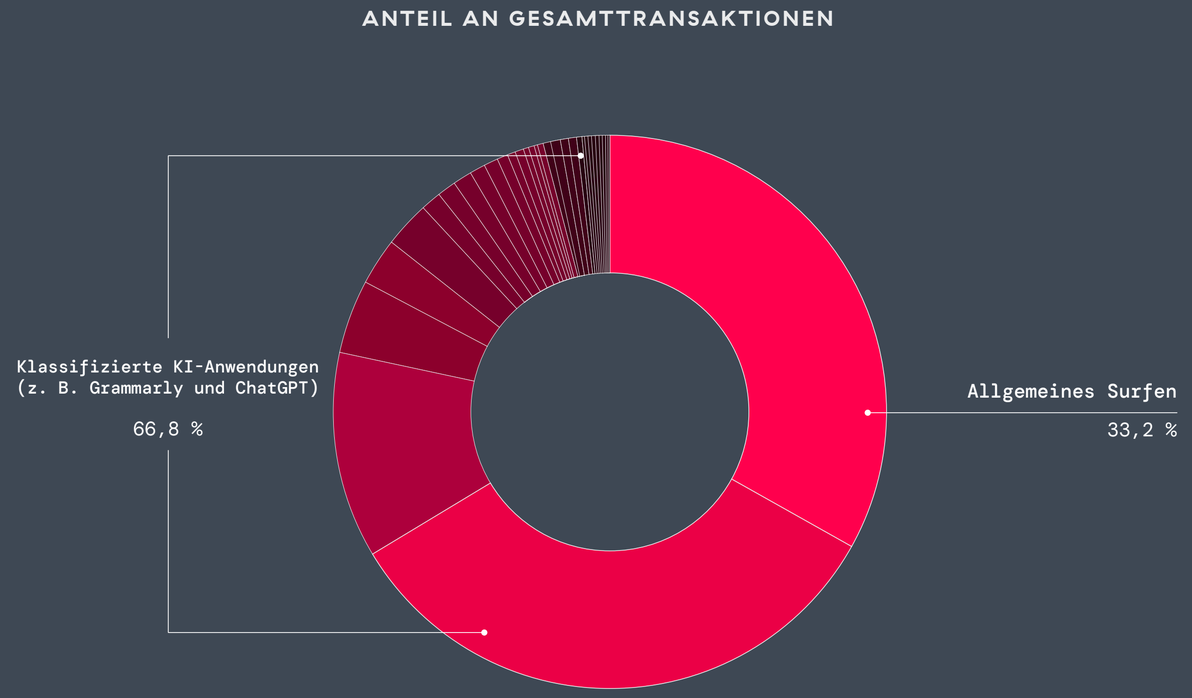


Abb. 2: Verteilung der KI-/ML-Transaktionen auf allgemeine und klassifizierte KI-Anwendungen

Führende LLM-Anbieter, Anwendungen und Abteilungen

Die Betrachtung der KI-Nutzung in Unternehmen nach LLM-Anbietern eröffnet einen besonderen Blick darauf, wie KI im großen Maßstab funktioniert. Mitarbeitende interagieren zwar täglich mit einzelnen Anwendungen und Funktionen, doch die Transaktionsmuster zeigen, welche Modellanbieter dahinterstehen. Erst der Blick auf die zugrunde liegenden LLM-Anbieter zeigt, wie sich KI-Nutzung in Unternehmen tatsächlich entwickelt.

Wichtigste Erkenntnisse zu LLM-Anbietern

- **OpenAI** war 2025 eindeutig der führende LLM-Anbieter. Mit 131 Milliarden Transaktionen übertraf das Unternehmen seinen stärksten Konkurrenten um mehr als das Dreifache. Die Einführung von GPT-5 im August weitete den Einsatz auf Programmierung, multimodales Denken und die Ausführung komplexer Aufgaben aus. Erweiterte Enterprise-API-Optionen mit stärkerem Datenschutz und Modellisolierung stärkten zudem OpenAIs Rolle als technisches Fundament für Copilots und KI-gestützte SaaS-Lösungen.

- **Codeium**, das 2025 in Windsurf umbenannt wurde, etablierte sich als zweitgrößte Quelle für LLM-Traffic in Unternehmen (42 Milliarden Transaktionen). Die starke Nutzung lässt sich wahrscheinlich auf seine speziell für Coding entwickelten proprietären Modelle zurückführen, die häufig in Entwicklungs-Pipelines und Engineering-Umgebungen eingesetzt werden. Das passt zu den späteren Ergebnissen der Abteilungsanalyse, nach denen Engineering den intensivsten KI-Einsatz zeigt.

- **Perplexity** landete im letzten Jahr mit 12 Milliarden Transaktionen auf Platz drei. Das Unternehmen bietet nicht nur KI-gestützte Suche, sondern betreibt auch eigene LLMs für seine Antwort-Engine. Entsprechend zeigt die Nutzung in Unternehmen, dass Unternehmen zunehmend auf KI-gestützte Recherche und Wissensaufbereitung setzen.

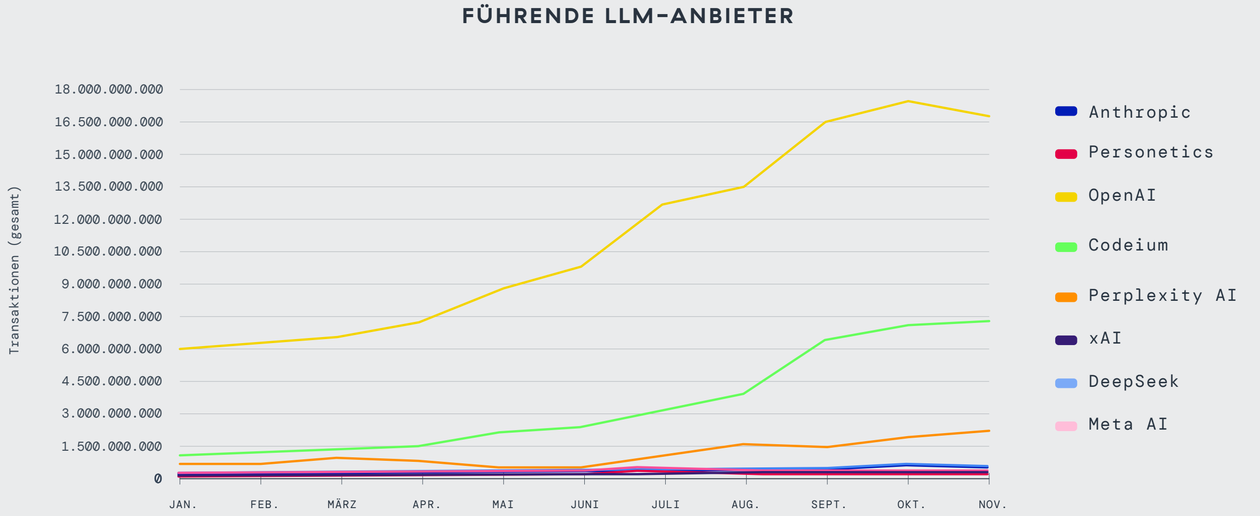


Abb. 3: Entwicklung der LLM-Transaktionen im Jahr 2025

Die meisten Transaktionen entfallen weiterhin auf eine Handvoll stark genutzter Anwendungen, die direkt in die täglichen Arbeitsprozesse integriert sind — Recherche, Bearbeitung, Texterstellung, Programmierung, Übersetzung und Zusammenarbeit.

Wichtigste Erkenntnisse zu Anwendungen

- **Grammarly** hat sich als aktivste KI-/ML-Anwendung im Unternehmensumfeld etabliert (38,7 % aller Transaktionen) und ChatGPT beim gesamten Transaktionsvolumen überholt. Angesichts von Funktionen, die von Textzusammenfassung bis hin zur gezielten Optimierung von Tonfall und Stil reichen, ist die prominente Rolle von Grammarly im Arbeitsalltag leicht nachzuvollziehen.
- **ChatGPT** behauptet sich mit 14,2 % weiterhin als dominierender Alltagshelfer. Unternehmen setzen das Tool flächendeckend für die Recherche oder das Erstellen von Texten ein, wodurch ChatGPT regelmäßig mit sensiblen Geschäftsdaten in Berührung kommt.
- **Codeium** gehört mit 5 % zu den fünf führenden Anwendungen. Das unterstreicht, dass KI mittlerweile ein fester Bestandteil der Softwareentwicklung ist, wobei Quellcode und proprietäre Logik routinemäßig verarbeitet werden.
- **DeepL** ist in globalen Unternehmen weiter auf dem Vormarsch (3,3 %). Die Lösung ist mittlerweile unverzichtbar geworden, um bei geschäftskritischen Inhalten über Sprachgrenzen hinweg den richtigen Ton zu treffen.
- **Microsoft Copilot** belegt den fünften Platz (3 %). Das Tool punktet vor allem durch die nahtlose Integration in Microsoft 365 und hilft dabei, alltägliche Produktivitätsaufgaben effizient zu automatisieren.

MEISTGENUTZTE KI-ANWENDUNGEN

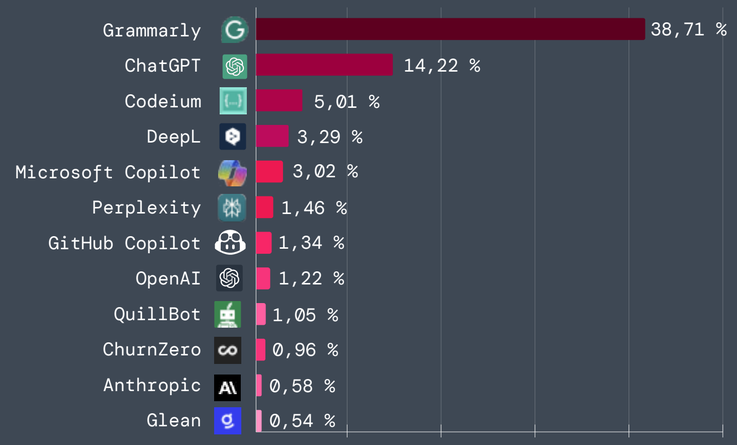


Abb. 4: Anteil führender KI-Anwendungen an den gesamten KI-/ML-Transaktionen

Hinweis: Die Zscaler Zero Trust Exchange erfasst ChatGPT-Transaktionen unabhängig von anderen OpenAI-Transaktionen.

DIE 20 MEISTGENUTZTEN KI-ANWENDUNGEN NACH TRANSAKTIONSVOLUMEN

Anwendung	Transaktionen (gesamt)
Grammarly	327.311.080.013
ChatGPT	120.227.890.252
Codeium	42.337.652.986
DeepL	27.847.680.087
Microsoft Copilot	25.503.137.940
Perplexity	12.386.054.978
GitHub Copilot	11.348.420.722
OpenAI	10.352.420.115
QuillBot	8.913.115.535
ChurnZero	8.153.526.358
Anthropic	4.922.983.385
Glean	4.542.501.122
GliaCloud	3.249.239.347
Claude	2.850.954.278
Google Gemini	2.604.461.019
SundaySky	2.483.835.170
Yellow Messenger	1.734.555.650
Cresta	1.585.454.178
Poe	1.483.703.558

Nach der Analyse der dominierenden KI-Tools folgt nun der nächste Schritt: der Blick auf die Teams.

ThreatLabz hat hierfür den KI-/ML-Traffic für eine Gruppe typischer Unternehmensabteilungen ausgewertet. So wird sichtbar, wie KI in der Praxis tatsächlich genutzt wird. In dieser Analyse wurden Anwendungen mit hoher Relevanz (mindestens eine Million Transaktionen) der jeweils primär nutzenden Abteilung zugeordnet. Wichtig: Die Prozentwerte zeigen die relative Nutzung innerhalb dieser Gruppe und spiegeln nicht den gesamten KI-Traffic des Unternehmens wider.

Zentrale Erkenntnisse nach Abteilungen

- Spitzenreiter bei der KI-Nutzung ist **Engineering** (48,9 %). Die Teams binden KI direkt in ihre täglichen Entwicklungszyklen ein. Der Vorteil: Schon minimale Zeitersparnisse zahlen sich durch die hohe Taktung der Releases massiv aus.

- Die **IT** belegt mit 31,8 % Platz zwei und ist in hohem Maße auf KI angewiesen. Hier steigert KI vor allem die Betriebseffizienz — etwa beim System-Support, der Fehlersuche oder der Automatisierung von Abläufen.

- An dritter Stelle steht die **Marketingabteilung** (6,9 %). Da sich der KI-Einsatz hier auf viele verschiedene Aufgaben im Bereich Content und Design verteilt, ist das Transaktionsvolumen zwar konstant, fällt aber niedriger aus als in den technischen Abteilungen

ANTEIL DER TRANSAKTIONEN NACH ABTEILUNG

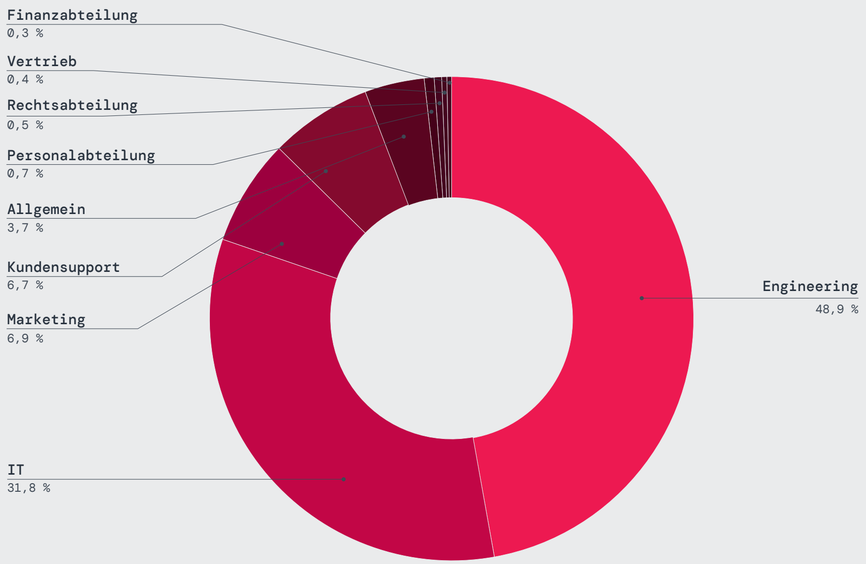


Abb. 5: Anteil der KI-/ML-Transaktionen nach zentralen Unternehmensabteilungen

Gesperrte Transaktionen

Bei der KI-Nutzung weht in Unternehmen seit 2025 ein strengerer Wind. Aus Sorge um Datensicherheit und Compliance wurden 39,2 % der KI-Transaktionen konsequent blockiert. Das zeigt deutlich: KI-Governance gehört im Sicherheitsalltag mittlerweile zum Standardrepertoire.

Die am stärksten von Kontrollmaßnahmen betroffenen Anwendungen gehörten gleichzeitig zu den am weitesten verbreiteten KI-Apps in Unternehmen. Grammarly verzeichnete den größten Anteil an blockierten Aktivitäten: 171,2 Milliarden Transaktionen wurden unterbunden, was 44,2 % aller blockierten KI-/ML-Transaktionen entspricht. Auch vielseitig einsetzbare KI-Lösungen stehen weiterhin unter strenger Beobachtung. So wurden ChatGPT und Microsoft Copilot mit 5,7 Milliarden bzw. 4,1 Milliarden blockierten Transaktionen häufig eingeschränkt, da der Zugriff auf unstrukturierte Daten das Risiko erhöht, dass sensible Unternehmensinformationen unbeabsichtigt geteilt werden.

Um firmeneigene Quellcodes und wertvolle Entwicklungsdaten zu schützen, haben Unternehmen auch bei KI-Coding-Tools wie Codeium oder Tabnine den Riegel vorgeschoben. Ebenso gerieten Anwendungen zur Text- und Sprachverarbeitung wie QuillBot und DeepL ins Visier der Sicherheitsverantwortlichen. Das Ziel ist klar: Unternehmen wollen verhindern, dass sensible Inhalte unkontrolliert an externe KI-Modelle abfließen.

AM HÄUFIGSTEN BLOCKIERTE KI-ANWENDUNGEN

1	Grammarly
2	GitHub Copilot
3	ChatGPT
4	Microsoft Copilot
5	QuillBot
6	Codeium
7	DeepL
8	Tabnine
9	Poe
10	Perplexity

An KI-Anwendungen übertragene Daten

Das Transaktionsvolumen allein lässt noch keine Rückschlüsse darauf zu, wie Unternehmen KI tatsächlich nutzen. Um ein vollständiges Bild zu erhalten, hat ThreatLabz zusätzlich die Datenmenge analysiert, die zwischen Unternehmensumgebungen und KI-/ML-Anwendungen übertragen wurde.

Im vergangenen Jahr ist der Datentransfer aus Unternehmen an KI-/ML-Anwendungen weiter gestiegen und erreichte 18.033 Terabyte (TB) — ein Zuwachs von 93 % im Vergleich zum Vorjahr. Den Löwenanteil dieser Datenbewegungen verursachten einige wenige beliebte Anwendungen. Ganz oben auf der Liste steht Grammarly mit 3.615 TB, gefolgt von ChatGPT (2.021 TB) und OpenAI (865 TB). Auch DeepL (625 TB) und Codeium (387 TB) gehören zu den Schwergewichten. All diese Anwendungen kommen in Bereichen zum Einsatz, in denen es um hochsensible Unternehmensdaten geht.

KI ist aus der täglichen Arbeit nicht mehr wegzudenken, wodurch das Volumen der übertragenen Unternehmensdaten stetig wächst. Die Analyse von sowohl Traffic- als auch Datenvolumen macht deutlich, wo die KI-Nutzung rasant zunimmt und wo Sicherheit und Governance am dringendsten erforderlich sind.

ANTEIL AM DATENTRANSFER

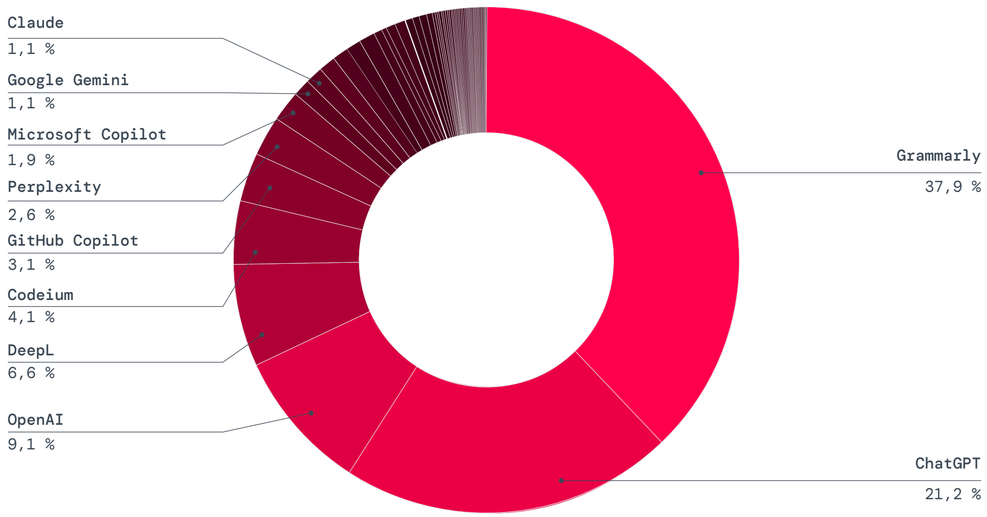
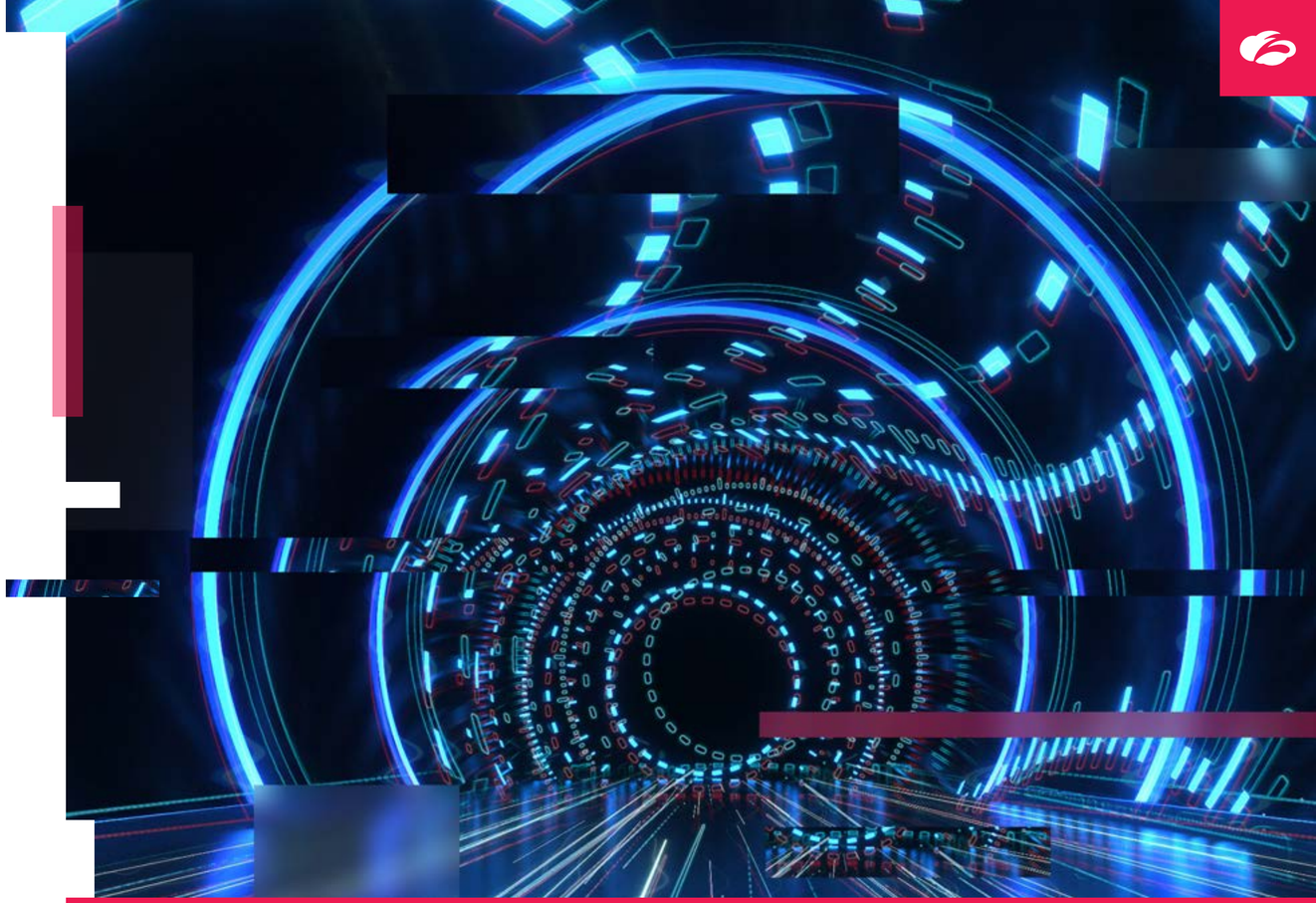


Abb. 6: Anteil führender KI-/ML-Anwendungen am gesamten Datentransfer



WICHTIGES ERGEBNIS

Insgesamt wurden **18.033 TB an Daten an KI-/ML-Anwendungen übertragen.** Das entspricht einer Steigerung von 93 % gegenüber dem Vorjahr.

Datenverlust durch KI-Anwendungen

Dass KI den Prozess von der ersten Idee bis zum fertigen Produkt extrem verkürzt, ist ein zweiseitiges Schwert: Was Zeit spart, gefährdet gleichzeitig den Datenschutz, da sensible Inhalte binnen *Sekunden* mit externen Modellen geteilt werden. Da KI-Funktionen zudem zunehmend in gängige SaaS-Anwendungen und -Services eingebettet sind, erfolgt die Datenübertragung oft automatisch. Dadurch steigt das Risiko, dass Daten unbemerkt nach außen abfließen.

Datenverluste an externe Modelle zu verhindern, hat sich zu einer der wichtigsten Sicherheitsprioritäten des Jahres entwickelt.

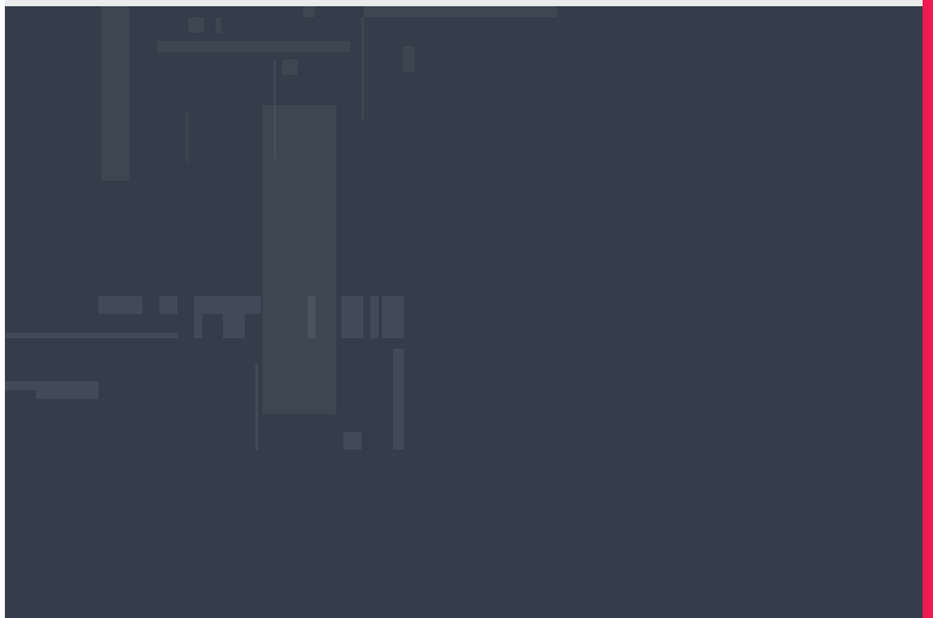
Innerhalb der Zscaler-Cloud sehen wir diesen Trend sehr deutlich an den steigenden Verstößen gegen KI-spezifische DLP-Richtlinien. Zscaler greift immer dann ein, wenn regulierte Inhalte wie Finanzdaten, personenbezogene Informationen, Quellcodes oder medizinische Daten über eine KI-App nach außen gelangen sollen. Ohne unsere KI-optimierte DLP-Lösung hätten Unternehmen keine Kontrolle darüber, welche Daten an externe Modelle übermittelt und dort potenziell offengelegt werden.

Gefahr droht vor allem dort, wo KI völlig gedankenlos eingesetzt wird — sei es bei Schreib-Tools, Coding-Assistenten oder KI-Features in gängigen Collaboration-Apps. Ihre einfache Handhabung ist ihr größtes Risiko. Diese Tools sehen sensible Inhalte zeitgleich mit den Usern, oft noch während sie erstellt werden.

Die Auswertung der Richtlinienverstöße zeigt deutlich, dass KI-Interaktionen sehr häufig einige der sensibelsten Daten eines Unternehmens betreffen.

KI-/ML-ANWENDUNGEN MIT DEN MEISTEN VERSTÖSSEN GEGEN DLP-RICHTLINIEN

Anwendung	Anzahl der DLP-Verstöße
ChatGPT	410.181.006
Codeium	242.263.311
GitHub Copilot	31.223.009
Claude	14.417.246
Wordtune	5.161.758
DeepL	2.037.613
QuillBot	1.960.391
Microsoft Copilot	1.858.952
Perplexity	1.235.129
Google Gemini	841.374



ChatGPT verzeichnete einen massiven Anstieg der DLP-Verstöße um 99,3 % im Vergleich zum Vorjahr. Dabei handelte es sich primär um das unbeabsichtigte Teilen von Klarnamen und Ausweisdaten, was auf eine Gefährdung von Kundendaten schließen lässt.

Noch deutlicher ist der **Anstieg bei Codeium:** Hier schossen die DLP-Verstöße um 100 % nach oben. Das signalisiert ein stark gewachsenes Risiko für den Verlust von wertvollem Quellcode und geistigem Eigentum.

Besonders alarmierend bei den KI-DLP-Verstößen ist das globale Ausmaß der gefährdeten Daten. Ob Ausweisdaten, Zahlungsinformationen, Quellcode oder medizinische Daten: Allesamt unterliegen sie strikten Regularien und dennoch geraten sie im Rahmen der KI-Nutzung zunehmend in unsichere Kanäle.

DIE 10 HÄUFIGSTEN VERSTÖSSE GEGEN KI-DLP-RICHTLINIEN

1	Klarnamen
2	Sozialversicherungsnummer (USA)
3	Unternehmensnummer (Japan)
4	NHS-Patientennummer (UK)
5	Quellcode
6	Medicare-Nummer (Australien)
7	National Provider Identifier (US-Gesundheitswesen)
8	Sozialversicherungsnummer (Kanada)
9	Medizinische Information
10	Kreditkartendaten

Die DLP-Trends bestätigen ein aus diversen Härtetests bekanntes Problem: KI-Systeme scheitern oft im ganz normalen Alltag, nicht erst bei komplexen Hackerangriffen. Details dazu finden Sie im Kapitel **Was die Unternehmens-KI wirklich gefährdet**.

Wie Sie den Abfluss von Daten durch GenAI-Apps verhindern, erfahren Sie unter **Sichere Einführung von GenAI-Apps in Unternehmen** weiter unten im Report.

Eingebettete KI auf dem Vormarsch

KI findet heute längst nicht mehr nur in Chatbots statt. Sie ist als eingebettete KI fest in unsere täglichen Programme integriert — oft als hilfreiche Assistenzfunktion für Zusammenfassungen, Empfehlungen oder Insights. Das Problem: Weil diese Features so selbstverständlich wirken, wird oft ignoriert, dass sie Zugriff auf Unternehmensdaten haben, ohne dass die gleichen Sicherheitsregeln greifen wie bei eigenständigen KI-Anwendungen. Das macht eingebettete KI zu einem wachsenden, aber fast unsichtbaren Sicherheitsrisiko.

Warum dieser Wandel wichtig ist? Eingebettete KI steigert die Effizienz, indem sie auf mehr Kontextdaten zugreift. Doch was der Produktivität dient, vergrößert gleichzeitig das Risiko eines Datenabflusses, sofern die Sicherheitsvorkehrungen nicht mit dieser Entwicklung mithalten. Wir haben die gängigsten Bedrohungsmuster zusammengefasst, die bei eingebetteten KI-Funktionen in Unternehmensanwendungen auftreten.

Wichtige Erkenntnisse

RISKANTE DATENFREIGABE DURCH ÜBERNOMMENE BERECHTIGUNGEN

Eingebettete KI greift in der Regel auf bestehende Zugriffskontrollen und Inhaltsberechtigungen zurück. Eingebettete KI nutzt die Rechte, die bereits im System hinterlegt sind. Das Problem: In vielen Firmen wurden Berechtigungen über Jahre unkontrolliert ausgeweitet. Die KI bringt nun Informationen ans Licht, die in vergessenen Gruppen oder geteilten Bereichen schlummern — und macht sie für Personen sichtbar, die diese für ihre Arbeit gar nicht brauchen. Was früher ein eher theoretisches Problem war, wird durch KI zu einer sehr präsenten Gefährdung sensibler Daten.

INDIREKTE PROMPT-MANIPULATION DURCH GESCHÄFTSINHALTE

Eingebettete KI liest im regulären Betrieb kontinuierlich Unternehmensinhalte wie E-Mails, Support-Tickets, Dokumentationen, Chat-Protokolle und Dateianhänge. Darin versteckte Anweisungen oder schädliche Inhalte könnten beeinflussen, wie die KI antwortet, was sie priorisiert oder wie sie Informationen darstellt. Wenn KI-Funktionen eng in Arbeitsabläufe integriert sind, können die Inhalte selbst zum Übertragungsweg für solche Manipulationen werden.

GEFÄHRDUNG DURCH DIE KI-LIEFERKETTE

Eingebettete KI stützt sich häufig auf eine Vielzahl von Komponenten. Sie nutzt ein Ökosystem aus Modellanbietern, Abfrageschichten, die Inhalte aus Unternehmenssystemen extrahieren, sowie Konnektoren für die Integration in SaaS-Anwendungen und Daten-Repositories. Jede dieser Schnittstellen erweitert die Angriffsfläche des Unternehmens. Da sich diese Funktionen ständig weiterentwickeln, kann sich das Risikoprofil durch Updates, Konfigurationsänderungen oder neu aktivierte Integrationen laufend verschieben.

RISIKEN DURCH AUTOMATISIERTE KI-WORKFLOWS

Der Sprung von der reinen Texterstellung zur aktiven Aufgabenführung durch KI birgt neue Gefahren. Wenn KI-Systeme eigenständig Prozesse anstoßen, Code schreiben oder Datenbanken pflegen, führen Fehler oder manipulierte Ergebnisse direkt zu betrieblichen Problemen. KI-Entscheidungen beeinflussen die gesamte nachfolgende Prozesskette oft so subtil, dass eine lückenlose Kontrolle und Nachverfolgung kaum noch möglich ist.

KI-EXPLOITS IN DER PRAXIS: DATENEXFILTRATION LEICHT GEMACHT

Wie gefährlich eingebettete KI sein kann, zeigen zwei prominente Beispiele aus dem Copilot-Ökosystem:

- **EchoLeak:** Ein „Zero-Click"-Angriff im Stil einer Prompt-Injection, der ohne Zutun des Users Daten stiehlt, indem er die automatisierte E-Mail-Analyse von Copilot ausnutzt.
- **Reprompt** ein „Single-Click"-Angriff, bei dem präparierte Prompts via URL-Parameter genutzt werden, um unerwünschtes Verhalten und Datenlecks auszulösen.

Da immer mehr SaaS-Anbieter KI standardmäßig in ihre Lösungen integrieren, müssen Unternehmen Transparenz, Governance und Datenschutz zwingend auf jene Anwendungen und Workflows ausweiten, in denen KI unsichtbar in die täglichen Arbeitsprozesse eingreift.

KI-/ML-Nutzung nach Branchen

Im Jahr 2025 hat KI in jedem Wirtschaftszweig Einzug gehalten. Jede Branche in der Zscaler-Cloud legte bei der KI-/ML-Nutzung zu. Doch nicht alle sind gleich weit. In manchen Sektoren ist KI bereits fest in die Wertschöpfung integriert, andere tasten sich noch an die Möglichkeiten heran.

Mit 23,3 % des KI-Traffics liegen Finanzdienstleister und Versicherungen erneut an der Spitze. Das ist wenig überraschend, da Daten und Automatisierung seit jeher ihr Kerngeschäft prägen. Den zweiten Platz belegt die **Fertigungsbranche** (19,5 %), die massiv auf KI für Smart Factories, Qualitätskontrolle und optimierte Lieferketten setzt. Besonders spannend ist die Entwicklung bei **Technologie und Kommunikation** sowie **Bildung** — hier sehen wir, wie unten detailliert aufgeführt, das schnellste Wachstum gegenüber dem Vorjahr.

ANTEIL AN KI-TRANSAKTIONEN NACH BRANCHE

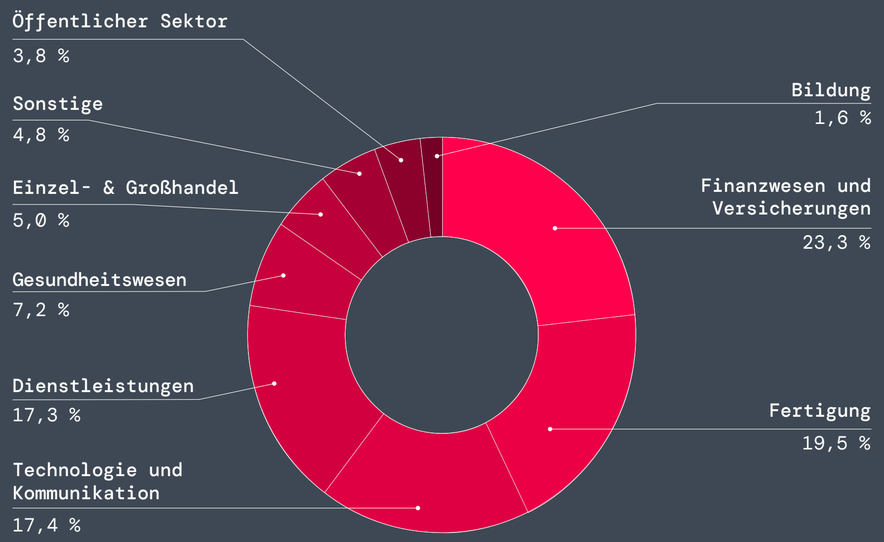


Abb. 7: Branchen mit dem größten Anteil an KI-Transaktionen

ANTEIL BLOCKIERTER KI-TRANSAKTIONEN NACH BRANCHE

Branche	% der KI-Transaktionen, die blockiert wurden
Finanzwesen und Versicherungen	39,1 %
Fertigung	22,1 %
Dienstleistungen	13,5 %
Gesundheitswesen	8,5 %
Technologie und Kommunikation	6,8 %
Öffentlicher Sektor	4,0 %
Sonstige	3,4 %
Einzel- & Großhandel	2,0 %
Bildung	0,6 %

Die Nutzung von KI findet nicht im luftleeren Raum statt; sie wird durch branchenspezifische Risiken, Compliance-Vorgaben und den Reifegrad der Sicherheitsprogramme beeinflusst.

Ein Blick auf die blockierten KI-Transaktionen verrät viel darüber, wie Sektoren das Spannungsfeld zwischen Innovation und Sicherheit meistern. Spitzenreiter bei der KI-Aktivität ist die Finanz- und Versicherungsbranche — doch sie blockiert gleichzeitig fast 40 % ihrer Anfragen. Diese Quote ist kein Zeichen von Ablehnung, sondern Ausdruck einer gelebten Sicherheitskultur in einem Sektor, in dem strikte Governance und Kontrolle bei neuen Technologien Pflicht sind.

Die Fertigungsindustrie, die beim KI-Transaktionsvolumen an zweiter Stelle steht, blockierte etwa 22 % ihres KI-Traffics. Dies deutet auf einen pragmatischen Mittelweg hin: Unternehmen in diesem Sektor nutzen die Vorteile von KI intensiv, setzen aber gleichzeitig auf starke Kontrollmechanismen. Ziel ist es, Datenlecks zu vermeiden und die Sicherheit zu gewährleisten — ein kritischer Faktor vor allem in sensiblen IoT- und OT-Umgebungen.

BRANCHEN-SPOTLIGHT

KI-Spitzenreiter Finanz- und Versicherungswesen: 230 Milliarden Transaktionen

Niemand nutzt KI so intensiv wie der Finanz- und Versicherungssektor — fast jede vierte KI-Transaktion in der Zscaler-Cloud entfällt auf diese Branche. Am beliebtesten sind Tools, die die tägliche Arbeit erleichtern: Grammarly, ChatGPT und Microsoft Copilot führen die Liste der Top-Apps das zweite Jahr in Folge an. Ob Zusammenfassungen von Analysen, Compliance-Berichte, Betrugserkennung oder die schnellere Bearbeitung von Versicherungsfällen — KI ist aus den Kernprozessen nicht mehr wegzudenken. Die Zahlen von Morgan Stanley[1] untermauern diesen Trend: Innerhalb weniger Monate stieg die KI-Nutzung bei Versicherungsunternehmen von 48 % auf stolze 71 %, bei Finanzdienstleistern von 66% auf 73 %.

Im Jahr 2025 haben verschiedene Markteinflüsse diese Beschleunigung noch einmal verstärkt. Banken sitzen der Modernisierungsstau und der Kostendruck im Nacken, weshalb sie bei der KI-Nutzung im Vergleich zu anderen Branchen schneller agieren. Gleichzeitig haben Versicherer mit teureren Schadensfällen und durch den Klimawandel verursachten Unsicherheiten zu kämpfen. Sie setzen daher verstärkt auf KI, um Preise präziser zu kalkulieren und Reaktionszeiten zu verkürzen.

Gleichzeitig agiert der Sektor beim Einsatz dieser Tools keineswegs sorglos. Finanzdienstleister und Versicherungsunternehmen blockierten über 39,1 % der KI-/ML-Transaktionen in der Zscaler-Cloud — ein klares Zeichen für das geschärfte Bewusstsein beim Datenschutz und den wachsamen Blick der Regulierungsbehörden. Es gilt, die Interaktionen von KI-Modellen mit hochsensiblen Finanzdaten streng zu reglementieren. Kurz gesagt: Die Branche gibt Vollgas, fährt aber auf Sicht und mit bremsbereitem Fuß.

Damit setzt sie auch 2026 die Maßstäbe für eine mutige, aber sichere KI-Transformation.

[1] Business Insider, **3 parts of the market where AI hype is turning into real returns, according to Morgan Stanley**, 24. Juli 2025.

BRANCHEN-SPOTLIGHT

Technologiesektor verzeichnet das schnellste Wachstum bei KI-Nutzung in Unternehmen: +202 % im Vorjahresvergleich

Kein anderer Sektor legte beim KI-/ML-Transaktionsvolumen in der Zscaler-Cloud im Jahr 2025 so stark zu wie die Technologiebranche (202,3 %). Dass Tech-Unternehmen bei generativer KI als Vorreiter der ersten Stunde gelten, ist bekannt. Doch der aktuelle Sprung verdeutlicht, wie massiv Softwareunternehmen, Cloud-Dienstleister, digitale Plattformen und Entwicklungsteams KI-Lösungen inzwischen vorantreiben — und zwar sowohl in ihren Produkten als auch in ihren eigenen Betriebsprozessen.

Führende Produktivitätsassistenten sind in Technologieunternehmen weit verbreitet. Sie unterstützen eine Vielzahl von Aufgaben, von der Codegenerierung über die technische Dokumentation bis hin zur Erstellung von Marketinginhalten. Dementsprechend gehörten Grammarly, Codeium, ChatGPT und Perplexity während unserer Analyse zu den am häufigsten genutzten KI-Anwendungen im Technologiesektor.

Trotz des rasanten Wachstums deckt KI in vielen Technologieunternehmen Lücken bei Transparenz und Richtliniendurchsetzung auf. Als Reaktion darauf investieren diese Unternehmen verstärkt in die Überwachung: Sie blockieren etwa 7 % der KI-Transaktionen — ein insgesamt zwar geringer Anteil, der jedoch deutlich über dem vieler anderer Branchen liegt. So verfeinern sie ihre Kontrollmechanismen, um eine sichere Bereitstellung zu gewährleisten.

BRANCHEN-SPOTLIGHT

Unauffällig, aber rasant: Bildungssektor steigert KI-Nutzung um 184 % im Jahresvergleich

Der Bildungssektor machte im Jahr 2025 zwar nur einen kleinen Teil der KI-/ML-Transaktionen in der Zscaler-Cloud aus, legte dafür aber ordentlich an Tempo zu. Mit fast 16 Milliarden KI-/ML-Transaktionen im Laufe des Jahres verzeichnete das Bildungswesen mit 184,4 % den zweithöchsten Zuwachs im Vorjahresvergleich. Damit vollzieht es eine der schnellsten KI-Transformationen im Vergleich aller untersuchten Branchen.

Dieser Anstieg deckt sich mit der zunehmenden Nutzung generativer KI in Lernprozessen und Unterrichtsabläufen. Anwendungen wie ChatGPT und Microsoft Copilot sind bei Studierenden, Schülern und Lehrkräften gleichermaßen beliebt — sei es als Schreibhilfe, zur Erstellung von Inhalten oder für die Unterrichtsplanung. Auch die Verwaltung nutzt KI, um Routineaufgaben effizienter zu gestalten. Das Spektrum reicht vom Entwerfen von Mitteilungen bis hin zur Verbesserung des Serviceangebots, was maßgeblich zum kontinuierlichen Anstieg des Transaktionsvolumens beiträgt.

Bemerkenswert ist, dass dieser Zuwachs nahezu reibungslos verlief. Weniger als 1 % der KI-/ML-Transaktionen im Bildungswesen wurden blockiert. Dies deutet darauf hin, dass der Großteil der Nutzung entweder ausdrücklich erlaubt ist oder in Umgebungen stattfindet, in denen Governance-Regeln und Sicherheitsvorgaben gerade erst entstehen. Verglichen mit größeren Sektoren zeigt sich das Bildungswesen daher verständlicherweise zurückhaltend. Schulen und Universitäten müssen zunächst Fragen zum Datenschutz und zur akademischen Integrität klären. Das ist vermutlich auch der Grund, warum die KI-Nutzung insgesamt noch hinter anderen Branchen zurückbleibt, obwohl die Kurve steil nach oben zeigt.

Dennoch bereitet das nahezu dreifache Wachstum innerhalb eines einzigen Jahres den Boden für strukturiertere und verantwortungsbewusstere KI-Initiativen sowie deren Integration im kommenden Jahr.

KI-/ML-Nutzung nach Land

Die geografische Verteilung der KI-/ML-Aktivitäten blieb 2025 weitgehend stabil, mit nur leichten Verschiebungen in den Randbereichen. Die **USA** sind als Epizentrum für die Entwicklung und den Einsatz von Unternehmens-KI fest etabliert und halten weiterhin den größten Anteil am KI-/ML-Traffic. Dennoch wuchs die KI-Nutzung in mehreren internationalen Märkten deutlich an.

Obwohl die USA bei der absoluten Nutzung weiterhin führen (218,9 Milliarden KI-/ML-Transaktionen, was 37,6 % der weltweiten Aktivität entspricht), nahm die KI-Einführung im Vorjahresvergleich in anderen Regionen schneller an Fahrt auf. Diese globale Beschleunigung zeigt sich am deutlichsten in **Indien**: Als zweitgrößte Quelle für KI-Aktivitäten in Unternehmen erreichte das Land 82,3 Milliarden Transaktionen — ein Plus von 309,9 % gegenüber dem Vorjahr. Indiens Wachstum deckt sich mit den staatlich geförderten Initiativen zur digitalen Transformation im Jahr 2025 sowie massiven Investitionen von öffentlicher und privater Seite in Technik und Know-how. Eine wachsende Zahl an KI-Fachkräften und Cloud-first-Architekturen, die eine schnelle und skalierbare Bereitstellung ermöglichen, haben vermutlich zu diesem im Vergleich zu den Vorjahren überproportionalen Wachstum beigetragen.

Abgesehen von den beiden Top-Nationen festigten auch andere etablierte Märkte den Trend: KI verbreitet sich in Unternehmen immer flächendeckender. **Kanada** verzeichnete 27,2 Milliarden Transaktionen (+229,9 % im Vorjahresvergleich) — unterstützt durch staatliche Investitionen in KI-Rechenkapazitäten und Programme zur Beschleunigung der Einführung in Unternehmen, insbesondere in regulierten Branchen. Das **Vereinigte Königreich** und **Japan** komplettierten die Top 5 mit Zuwächsen von 117,5 % beziehungsweise 122,8 %.

Diese breite geografische Präsenz unterstreicht, dass KI längst zum Standardrepertoire in Unternehmen gehört. Für Sicherheitsteams bedeutet das: Sie müssen dieser global verteilten Nutzung Rechnung tragen und eine lückenlose Überwachung über alle Standorte und Ländergrenzen hinweg sicherstellen.

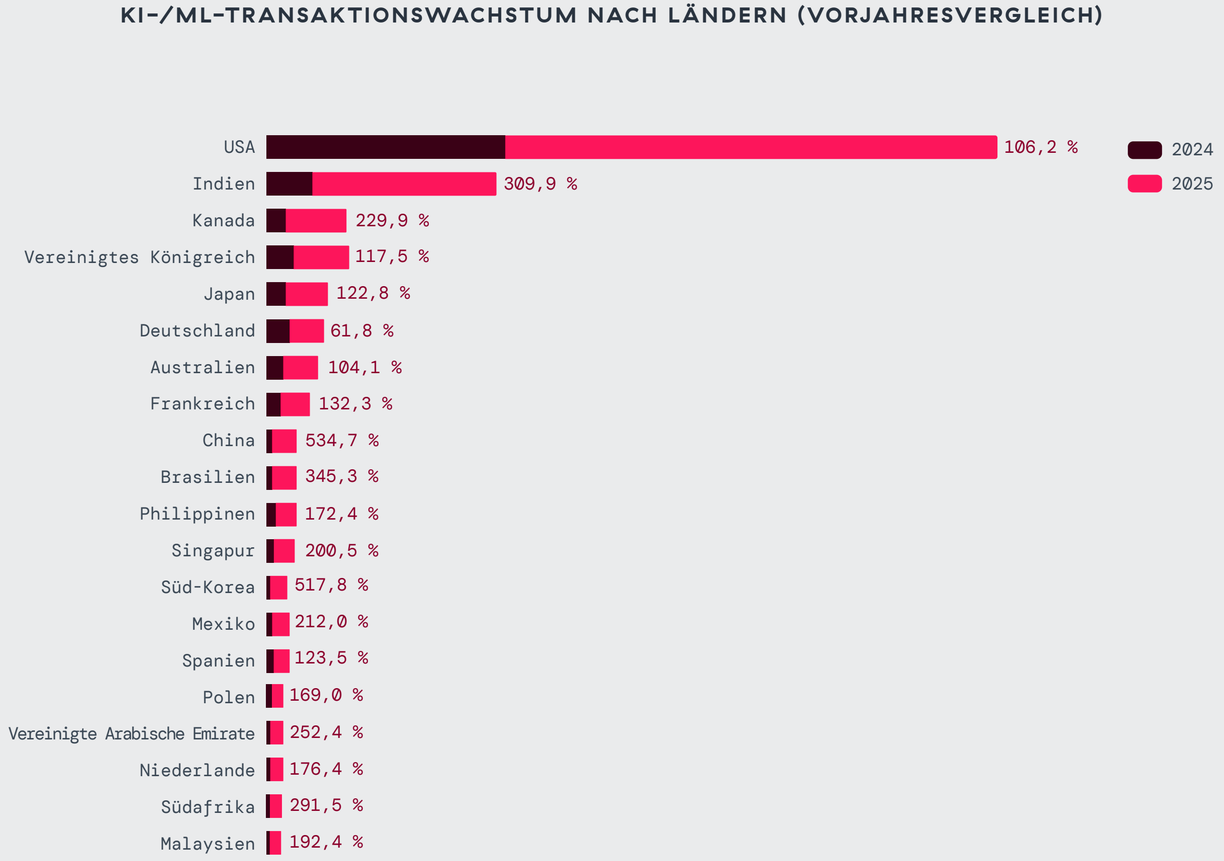


Abb. 8: KI-/ML-Transaktionswachstum im Vorjahresvergleich nach Ländern (Top 20 nach Transaktionsvolumen)

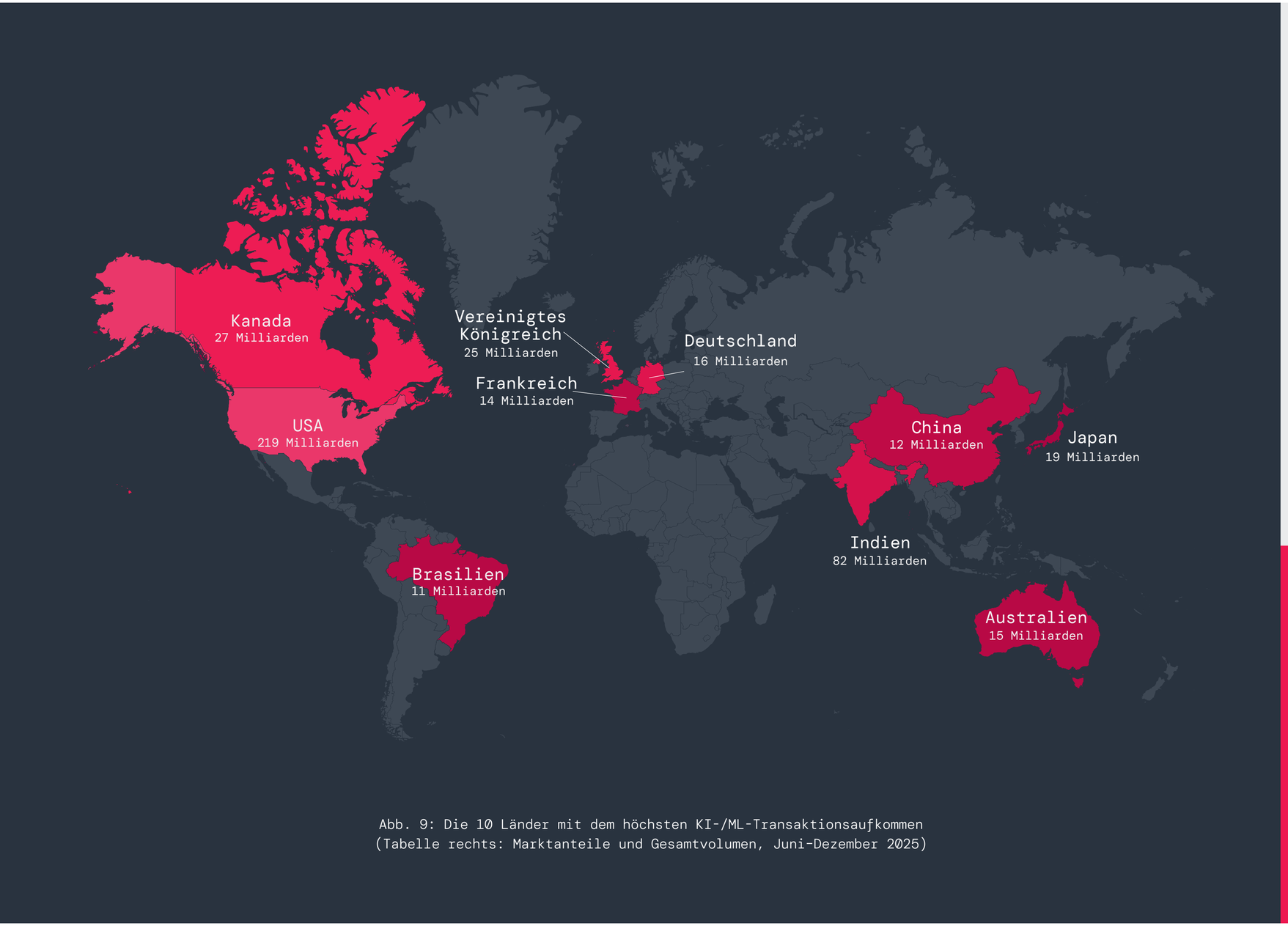


Abb. 9: Die 10 Länder mit dem höchsten KI-/ML-Transaktionsaufkommen (Tabelle rechts: Marktanteile und Gesamtvolumen, Juni-Dezember 2025)

Land	Marktanteil	KI-/ML-Transaktionen
USA	37,6 %	219 Mrd.
Indien	14,1 %	82 Mrd.
Kanada	4,7 %	27 Mrd.
Vereinigtes Königreich	4,3 %	25 Mrd.
Japan	3,2 %	19 Mrd.
Deutschland	2,7 %	16 Mrd.
Australien	2,6 %	15 Mrd.
Frankreich	2,4 %	14 Mrd.
China	2,0 %	12 Mrd.
Brasilien	1,8 %	11 Mrd.

REGIONALER SNAPSHOT

Einblicke in die Region EMEA

Die KI-/ML-Aktivitäten in der EMEA-Region konzentrierten sich weiterhin auf eine kleine Anzahl etablierter europäischer Märkte. Auf das Vereinigte Königreich, Deutschland, Frankreich und Spanien entfiel fast die Hälfte aller regionalen Transaktionen. Während das Vereinigte Königreich global gesehen eine kleinere Rolle spielt, dominiert es den EMEA-Raum deutlich. Mit einem Anteil von 20,3 % am KI-/ML-Traffic war das Land zwischen Juni und Dezember 2025 der unangefochtene Spitzenreiter der Region.

Deutschland folgte mit 12,5 % der Transaktionen im EMEA-Raum. Treiber dieser Entwicklung war die fortschreitende KI-Integration in der Fertigungsbranche, die mehr als 5,5 Milliarden KI-/ML-Transaktionen generierte. Dicht dahinter lag Frankreich mit einem Anteil von 11 % am regionalen Aufkommen. Gestützt wurde dies durch staatliche Initiativen wie die Strategie „France 2030“, die massive Investitionszusagen für KI umfasst, sowie durch die Rolle Frankreichs als Gastgeber des internationalen KI-Aktionsgipfels.

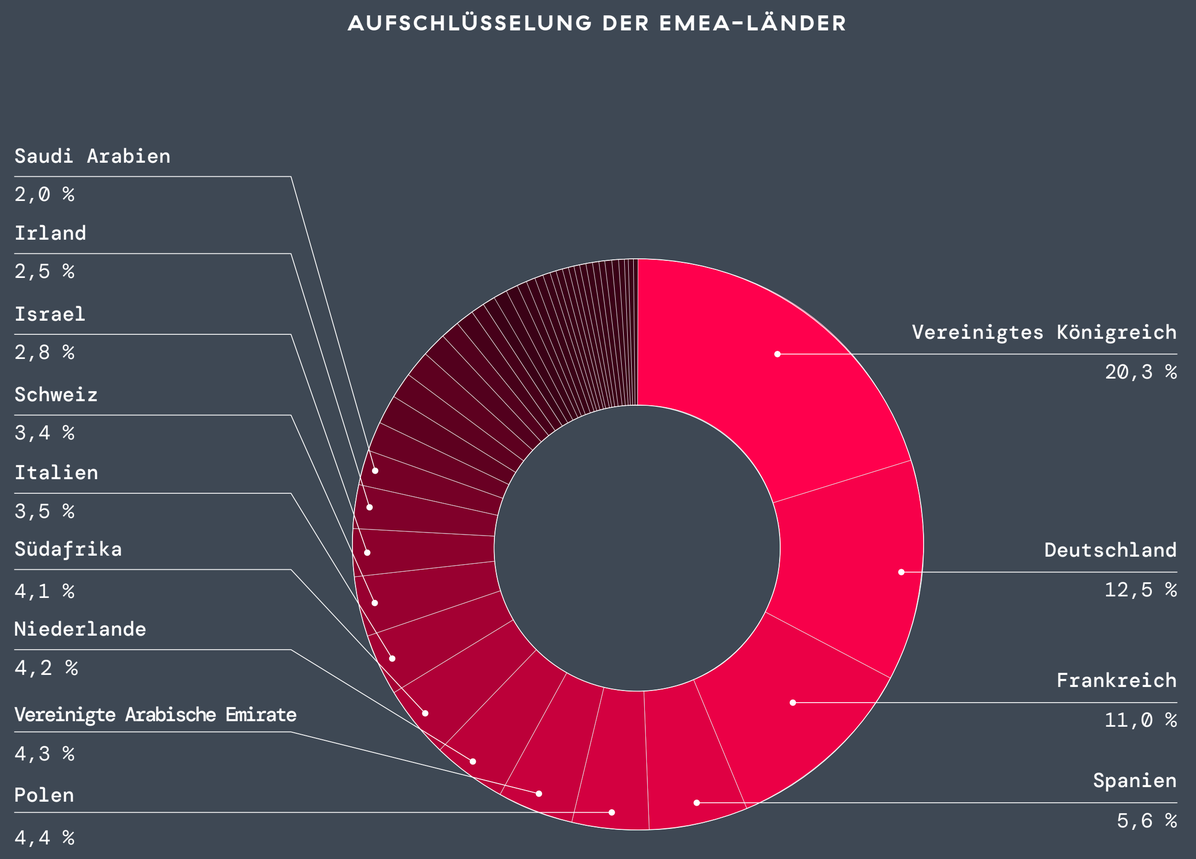


Abb. 10: Anteil der KI-Transaktionen nach Ländern in der Region EMEA

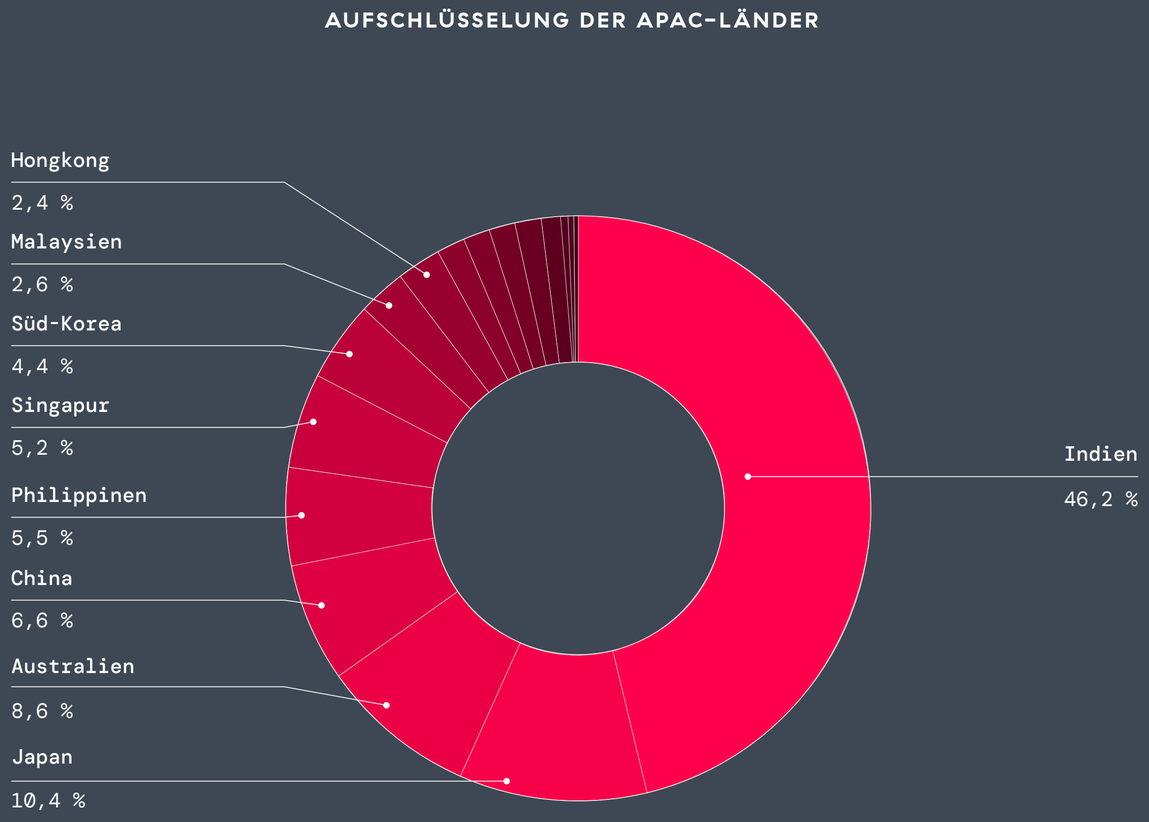


Abb. 11: Anteil der KI-Transaktionen nach Ländern in der Region APAC

REGIONALER SNAPSHOT

Einblicke in die Region Asien-Pazifik

In der APAC-Region (Asien-Pazifik) zeigt sich bei der KI-Nutzung eine deutliche Schieflage: Einem sehr wachstumsstarken Markt stehen mehrere gefestigte Wirtschaftsnationen gegenüber. Indien, Japan und Australien machten zusammen den Großteil der regionalen Transaktionen aus. Indien gibt dabei klar das Tempo vor und ist für fast die Hälfte (46,2 %) des Traffics in APAC verantwortlich. Haupttreiber ist hier der Technologie- und Kommunikationssektor, der 31 Milliarden Transaktionen beisteuerte.

Japan folgte mit 10,4 % der Transaktionen vor dem Hintergrund einer sich wandelnden nationalen KI-Politik. Die japanische Regierung verabschiedete ein Gesetz zur Förderung von KI, das den Einsatz der Technologie in Industrie und Unternehmen durch koordinierte Leitlinien unterstützt. Australien kommt auf einen Anteil von 8,6 %. Das Land legt dabei seit jeher großen Wert darauf, KI-Lösungen nicht nur schnell, sondern vor allem verantwortungsbewusst und sicher einzuführen.

KI-Risiken und -Bedrohungslandschaft in Unternehmen

Wie unsere Untersuchungen belegen, durchzieht KI mittlerweile jede Ebene des Unternehmens — von öffentlichen GenAI-Tools über interne LLMs bis hin zu KI-gestützten SaaS-Suites. Mit der zunehmenden Nutzung müssen Organisationen eine breitere und komplexere Angriffsfläche absichern. Die bedeutendsten Risiken lassen sich in die folgenden Kategorien unterteilen.

Datenexposition und Abfluss sensibler Informationen

KI-Systeme verarbeiten einige der sensibelsten Daten eines Unternehmens — von Quellcode und Kundendaten bis hin zu Finanzinformationen und rechtlichen Dokumenten — oft ohne ausreichende Sicherheitsvorkehrungen. Diese Exposition resultiert häufig aus der Nutzung von Schatten-KI in öffentlichen Tools wie ChatGPT, Grok und DeepSeek sowie aus SaaS-KI mit übermäßigen Berechtigungen wie Microsoft Copilot, die aufgrund von Fehlkonfigurationen oder ungenauen Kennzeichnungen Daten offenlegt. Parallel dazu können unkontrollierte RAG-Pipelines (Retrieval-Augmented Generation) unbemerkt regulierte Daten in private Modelle einspeisen. Sobald sensible Informationen an ein KI-System übermittelt werden, können sie gespeichert, wiederverwendet oder sogar durch Prompt-Manipulation oder das Modellverhalten preisgegeben werden, was die tägliche KI-Nutzung in ein echtes Datenrisiko verwandelt.

Mangelnde Transparenz über KI-Nutzung und User-Prompts

Viele Unternehmen haben nach wie vor Schwierigkeiten, grundlegende Fragen zur täglichen KI-Nutzung zu beantworten. Den Sicherheitsteams fehlt oft der klare Überblick darüber, welche KI-Tools die Mitarbeiter verwenden, welche Prompts sie übermitteln und ob dabei sensible Daten gefährdet sind. Zudem ist nicht immer ersichtlich, welche Teams generative KI bereits für kritische Workflows nutzen. Werden Prompts nachträglich überprüft, offenbaren sie häufig Versuche von Prompt-Injection, Manipulationsmuster oder nicht richtlinienkonformes Verhalten, das Sicherheitsvorkehrungen mit minimalem Aufwand umgeht. Da die meisten Unternehmen jedoch nicht über die Tools verfügen, um diese Aktivitäten in Echtzeit zu überwachen, bleibt die KI-Governance oft reaktiv. Sie greift erst ein, wenn bereits ein Problem aufgetreten ist.

Datenqualität, Halluzinationen und Modell-Manipulation

Wenn KI-Systeme Fehler machen, hat das im Geschäftsleben echte Folgen. 2025 häuften sich Fälle von sogenannten Halluzinationen: Die KI lieferte falsche Informationen mit absoluter Überzeugung. Selbst moderne RAG-Systeme blieben davon nicht verschont, wenn sie mit qualitativ schlechten oder einseitigen Daten gefüttert wurden. Gefährlich wird es bei gezielten Angriffen, wie **Red-Teaming-Übungen und Praxistests** zeigten: Hacker können die Informationsquellen der KI manipulieren (Data Poisoning) oder durch minimale Änderungen an den Prompts die Sicherheitsmechanismen aushebeln. Solche Fehler bei der Faktenprüfung und Logik erschüttern das Vertrauen in die Technologie. Werden diese Ausgaben ungeprüft übernommen, drohen Fehlentscheidungen und unkalkulierbare Risiken.

Nicht erfasste und ungeschützte private KI-Modelle

Unternehmen setzen heute eine Mischung aus verwalteten und nicht verwalteten Modellen ein sowie KI-Funktionen, die nativ in Plattformen wie Salesforce, ServiceNow und Atlassian eingebettet sind.

Dennoch mangelt es vielen Organisationen nach wie vor an:

- einem vollständigen Inventar aller genutzten Modelle und Services,
- Klarheit darüber, mit welchen Daten das jeweilige Modell in Berührung kommt,
- einer Validierung der Modellsicherheit, der Patch-Stände oder des Schwachstellenstatus,
- einer Governance für Quellcode-Repositories, die KI-Workflows speisen.

Diese mangelnde Übersicht wird besonders gefährlich, wenn private Modelle dieselben Schwachstellen für Prompt-Injection, RAG-Poisoning und Datenlecks aufweisen wie öffentliche Systeme. Wenn Modelle und deren Datenflüsse unbekannt sind, können Unternehmen weder Richtlinien durchsetzen noch Risiken sinnvoll bewerten.

Datenschutzrisiken durch unterschiedliche Anbieterstandards

KI-Anbieter verfolgen unterschiedliche Ansätze beim Umgang mit Unternehmensdaten. Prompts werden teils gespeichert, für Trainingszwecke wiederverwendet oder auf eine Weise protokolliert, die nicht immer transparent ist. Auch Zugriffskontrollen und die Data-Lineage der Modelle variieren stark von Anbieter zu Anbieter. Diese Inkonsistenz führt zu Compliance-Problemen im Rahmen von Frameworks wie DSGVO, HIPAA und PCI DSS. Das Risiko verschärft sich dadurch, dass SaaS-Anwendungen standardmäßig aktivierte KI-Funktionen enthalten, die etablierte Genehmigungsprozesse umgehen und dazu führen, dass Unternehmensrichtlinien nicht mehr mit regulatorischen Vorgaben übereinstimmen.

Reale Bedrohungen und Schwachstellen

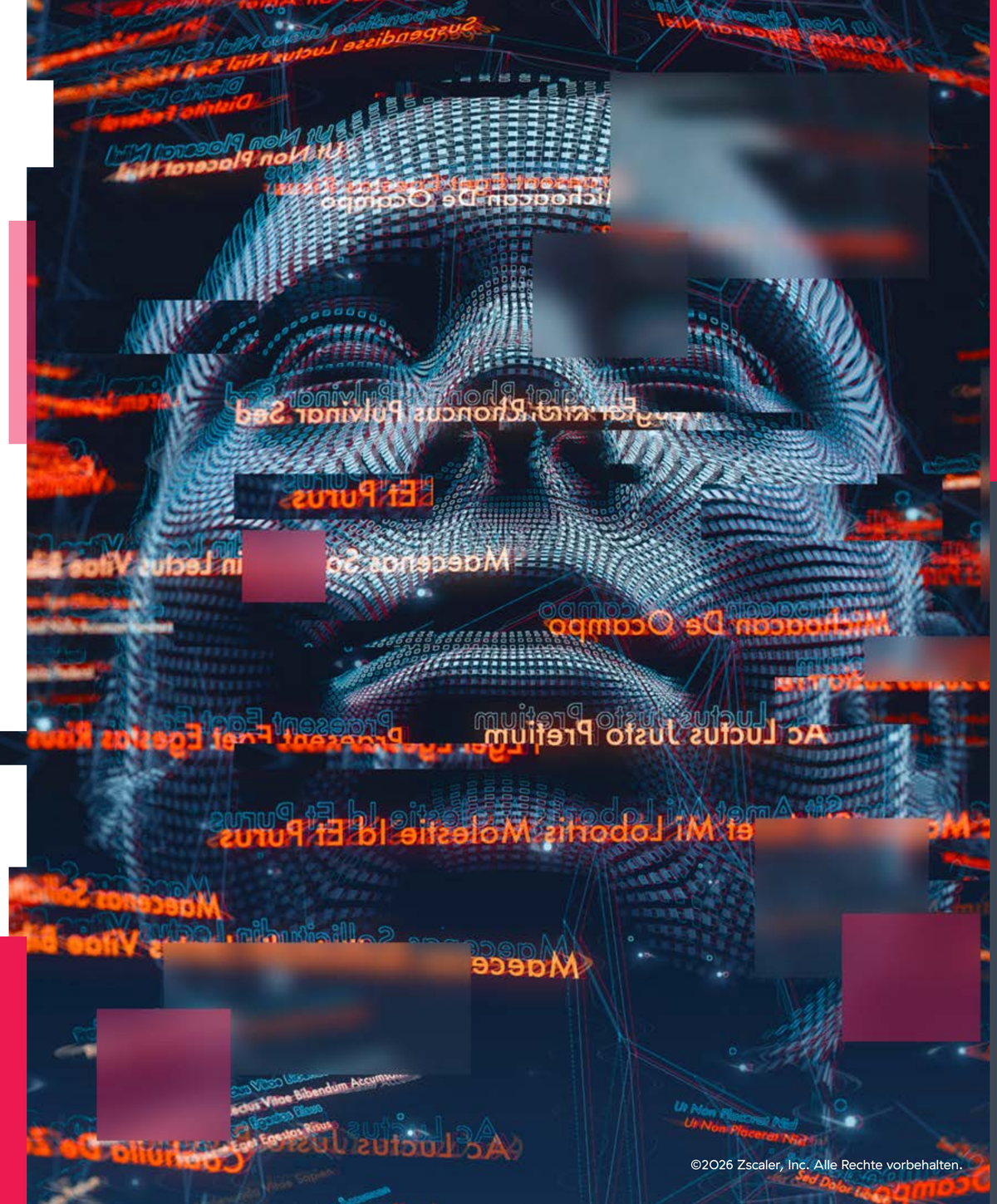
Die Kernrisiken bei der Einführung von KI in Unternehmen manifestierten sich im Jahr 2025 weiterhin in der Praxis. Bedenken hinsichtlich Datenexposition, mangelnder Transparenz bei der KI-Nutzung, Halluzinationen und weiteren Faktoren traten als konkrete Sicherheitsbedrohungen und operative Schwachstellen in Unternehmensumgebungen zutage. Reale Vorfälle und Testergebnisse verdeutlichten, dass diese Risiken direkt daraus resultieren, wie KI-Systeme implementiert, mit Daten vernetzt und in tägliche Arbeitsabläufe integriert werden.

Einige der bedeutendsten zugrunde liegenden Risiken zeigten sich in Form von KI-gestütztem Social Engineering, Datenlecks in KI-Anwendungen und -Assistenten sowie in ersten Missbrauchsfällen von agentenbasierten und teilautonomen KI-Systemen.

Social Engineering erreichte durch KI eine neue Eskalationsstufe: Angreifer nutzen generative KI, um Identitäten noch glaubwürdiger zu fälschen. Deepfakes in Form von Audio- und Video-Phishing (Vishing) wurden 2025 zu einer realen Gefahr. Behörden, unter anderem aus den USA, warnten vor Vorfällen, bei denen Kriminelle Regierungsvertreter mit KI-generierten Stimmen und Nachrichten imitierten.[2] Angreifer nutzen KI heute für maßgeschneiderte Deepfakes, die exakt auf die jeweilige Zielperson und deren Befugnisse im Unternehmen abgestimmt sind.

2025 markierte den ersten ernsthaften Fall von **agentischer KI in der Cyber-Spionage**. Eine staatlich geförderte Gruppe aus China automatisierte 80—90 % der gesamten Angriffskette mittels KI-Agenten — einschließlich Aufklärung, Exploit-Validierung, Diebstahl von Zugangsdaten, lateraler Bewegung und Datenexfiltration. Menschen griffen nur noch strategisch ein. Dieser Vorfall demonstrierte, wie klassische Angriffsmuster nun in Maschinengeschwindigkeit ausgeführt werden können. Für die IT-Abwehr bedeutet das einen Paradigmenwechsel, da herkömmliche Erkennungs- und Reaktionsmaßnahmen gegen dieses Tempo kaum noch ausreichen.

Über den direkten Missbrauch von KI-Systemen hinaus haben Angreifer begonnen, KI in ihre eigenen Entwicklungsprozesse zu integrieren. In mehreren von ThreatLabz beobachteten Kampagnen wies Malware Merkmale auf, die auf eine KI-gestützte Codegenerierung schließen lassen. Ein klares Indiz dafür, dass GenAI immer häufiger genutzt wird, um Schadsoftware effizienter und schneller zu entwickeln.



Die folgenden Fallstudien belegen die KI-Risiken anhand konkreter Beispiele — von KI-gestützten Täuschungsmanövern und Angriffsszenarien bis hin zu Red-Teaming-Tests, die aufzeigen, wie sich KI-Systeme in Unternehmen unter realen Angriffsbedingungen verhalten.

[2] Cybersecurity Dive, **FBI warns senior US officials are being impersonated using texts, AI-based voice cloning**, 16. Mai 2025.

FALLSTUDIE

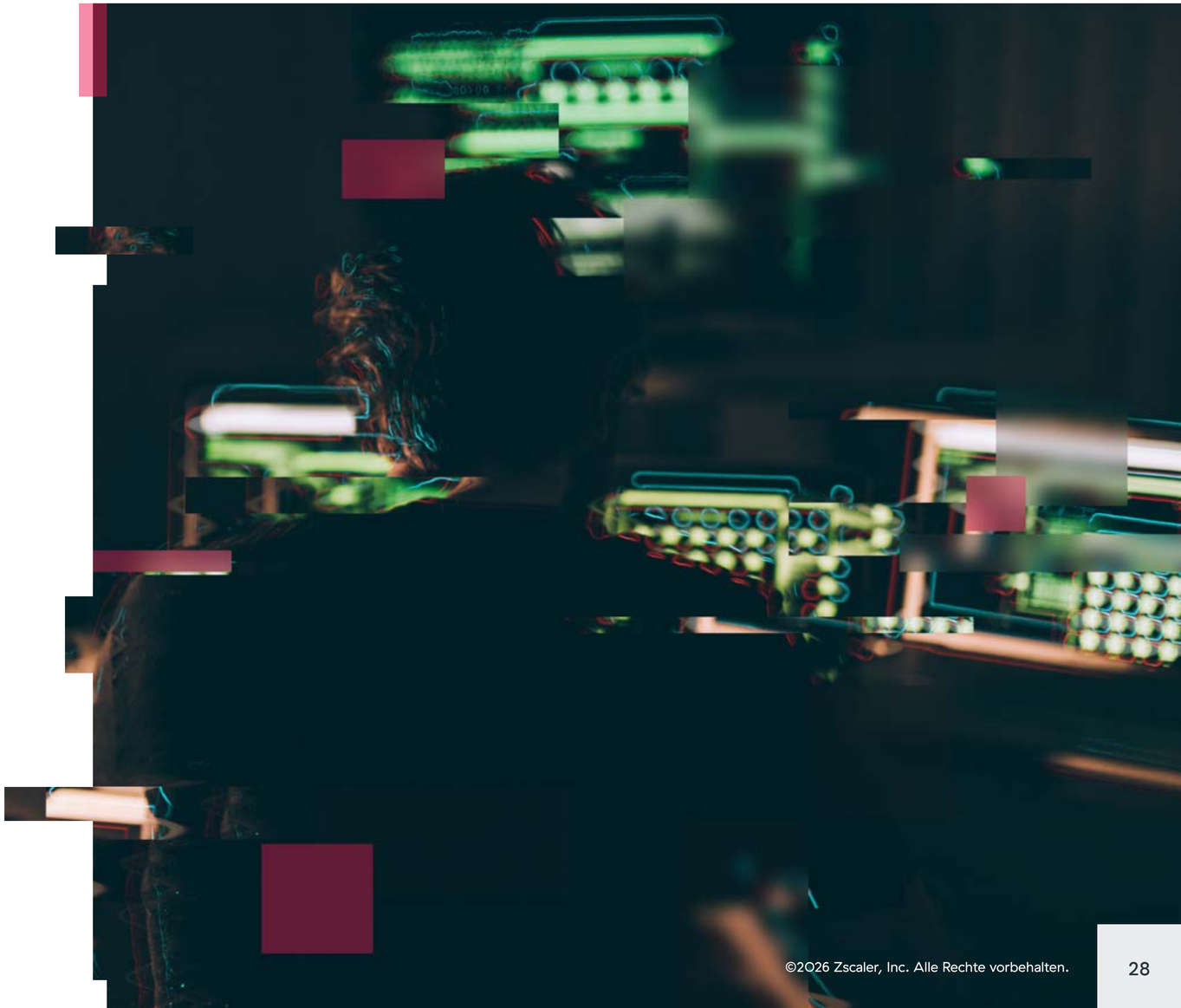
GenAI-gestützte Malware und Social Engineering in Kampagnen mit Verbindungen zu Nordkorea

Die vorliegende Fallstudie zeigt: GenAI verändert zwar nicht die Ziele der Hacker, macht ihre Angriffe aber deutlich schlagkräftiger.

In der **Kampagne „Contagious Interview“**, die Aktivitäten mit Bezug zu Nordkorea und dem weiteren Umfeld nordkoreanischer IT-Kräfte zugeordnet wird, beobachtete ThreatLabz, wie Akteure generative KI einsetzten, um Social Engineering im großen Maßstab zu betreiben. Dabei wurden überzeugende Fake-Identitäten erstellt und operationalisiert sowie KI-gestützte Codierung bei der Malware-Entwicklung genutzt. KI sorgt dafür, dass sowohl das Eindringen der Angreifer als auch deren darauf folgenden Aktivitäten immer schwerer von legitimen Vorgängen zu unterscheiden sind, was die Messlatte für Erkennung und Reaktion deutlich höher legt.

Ressourcenentwicklung & Social Engineering (Täuschung im Bewerbungsprozess)

Die Kampagne beginnt mit der Erstellung digitaler Identitäten mittels GenAI-Technologie. Dabei werden umfassende Leitfäden erstellt und professionelle, aber nicht rückverfolgbare Profilbilder generiert. In Remote-Vorstellungsgesprächen werden Deepfakes und Stimmverzerrer eingesetzt, damit die Angreifer anonym bleiben. Ziel dieses Betrugs ist es, Sicherheitschecks zu umgehen und sensible technische Positionen zu besetzen.



Die folgenden Ergebnisse unterstreichen, wie massiv sich die Vorbereitungsphase dieser Operation auf KI stützt.

INTERVIEW-TRAINING PER KI: LERNHILFEN FÜR HACKER

Angreifer nutzen generative KI, um umfassende Skripte für technische Interviews zu entwerfen.

Beispiel: Ein 70-seitiges Dokument bereitet die Akteure gezielt auf Fachfragen zu Backend-Systemen und Web3 vor.

Wesentliche Indikatoren für KI-Nutzung:

- Die Antworten in den Leitfäden enthalten typische Standardformulierungen wie „Certainly!“ („Sicher!“/„Gerne!“) (Abbildung 12).
- Überreste von Markdown-Formatierungen deuten stark darauf hin, dass die Inhalte per Copy-and-paste direkt aus der Ausgabe des KI-Modells übernommen wurden (Abbildung 13).

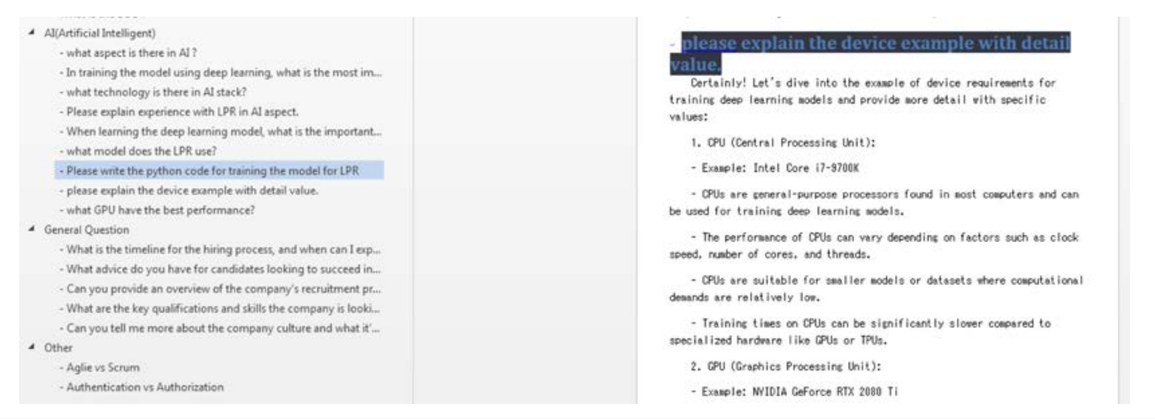
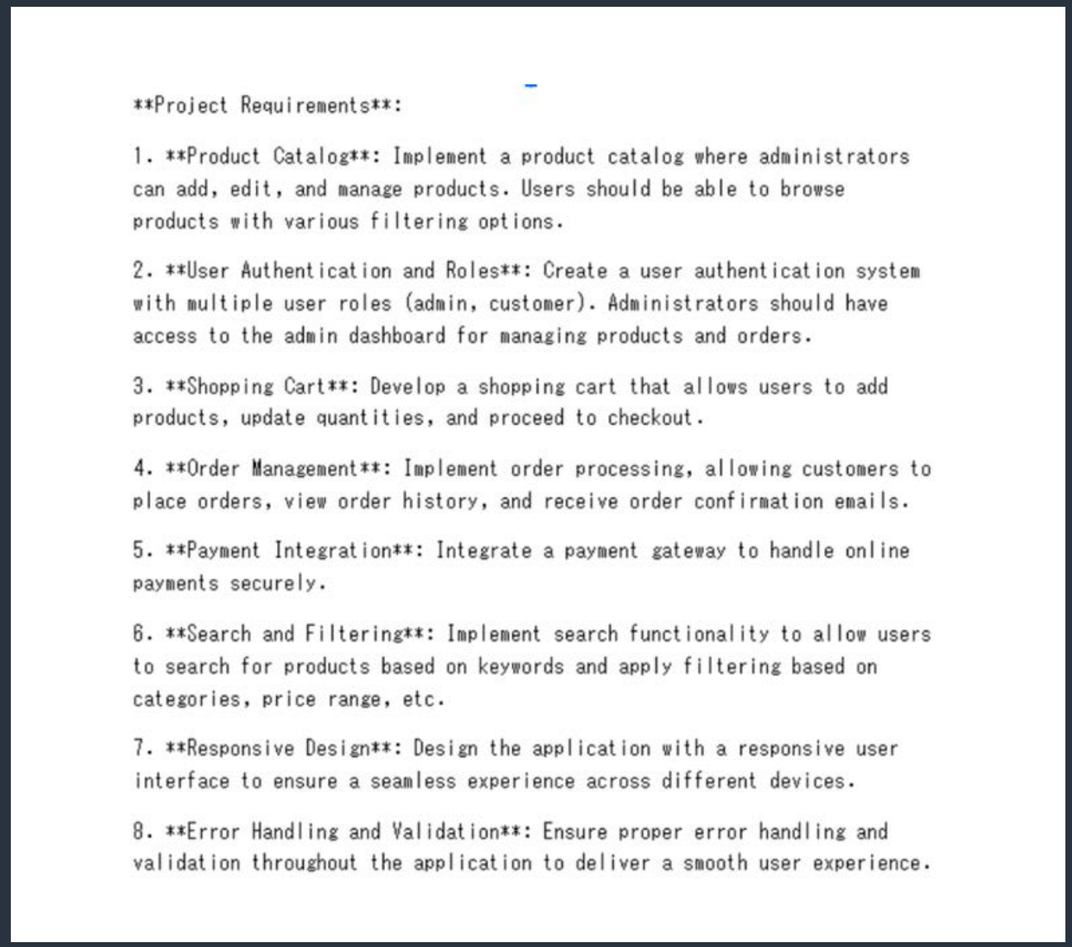


Abbildung 12: Fragen-und-Antworten-Katalog aus dem Leitfaden, der typische sprachliche Merkmale generativer KI aufweist.



Project Requirements:

1. **Product Catalog**: Implement a product catalog where administrators can add, edit, and manage products. Users should be able to browse products with various filtering options.

2. **User Authentication and Roles**: Create a user authentication system with multiple user roles (admin, customer). Administrators should have access to the admin dashboard for managing products and orders.

3. **Shopping Cart**: Develop a shopping cart that allows users to add products, update quantities, and proceed to checkout.

4. **Order Management**: Implement order processing, allowing customers to place orders, view order history, and receive order confirmation emails.

5. **Payment Integration**: Integrate a payment gateway to handle online payments securely.

6. **Search and Filtering**: Implement search functionality to allow users to search for products based on keywords and apply filtering based on categories, price range, etc.

7. **Responsive Design**: Design the application with a responsive user interface to ensure a seamless experience across different devices.

8. **Error Handling and Validation**: Ensure proper error handling and validation throughout the application to deliver a smooth user experience.

Abb. 13: Markdown-Formatierung, die darauf hindeutet, dass der Inhalt wahrscheinlich direkt aus einer GenAI-Ausgabe kopiert wurde.

Fallstudie: GenAI-gestützte Malware und Social Engineering in Kampagnen mit Verbindungen zu Nordkorea

IDENTITÄTSFÄLSCHUNG MITHILFE KI-GESTÜTZTER BILDBEARBEITUNG

Nordkoreanische IT-Mitarbeiter nutzen KI-Bildgenerierung und Bearbeitungstools, um gefälschte digitale Identitäten für Lebensläufe, eigene Portfolio-Webseiten und GitHub-Profile zu erstellen.

Beispiel: Durch KI verbesserte Porträtfotos sollen Kompetenz ausstrahlen und sind oft an westliche Schönheitsideale angepasst. Zudem wird das reale Arbeitsumfeld durch manipulierte Hintergründe unkenntlich gemacht.

Wesentliche Indikatoren für KI-Nutzung:

- Unnatürliche Perfektion: Die Porträts wirken durch die KI-Optimierung oft zu glatt und künstlich (Abbildung 14).
- Bearbeitungsspuren: In den Metadaten oder an den Bildkanten lassen sich Artefakte finden, die typisch für eine automatisierte Hintergrundentfernung sind (Abbildung 15).

Abb. 14: Originalbild (links) und mit KI bearbeitete Bilder (rechts)

Abb. 15: KI-optimiertes Profilbild

Fallstudie: GenAI-gestützte Malware und Social Engineering in Kampagnen mit Verbindungen zu Nordkorea

Erstzugriff durch manipulierte Software

Nach der erfolgreichen Kontaktaufnahme attackieren die Hacker gezielt Personen in Schlüsselpositionen, etwa Ingenieure im Krypto-Bereich. Durch gezielte Manipulation werden diese dazu gebracht, Schadsoftware zu installieren, die als nützliches Tool getarnt ist. Ein typisches Beispiel sind manipulierte NPM-Pakete: Die Angreifer verstecken ihren Schadcode in vermeintlich echten Entwickler-Ressourcen, um sich dauerhaft im Netzwerk des Opfers festzusetzen.

Bei der Analyse stießen wir auf einen entscheidenden Punkt: Mehrere der bösartigen Skripte zeigten deutliche Anzeichen für eine KI-basierte Generierung. Abbildung 16 verdeutlicht dies. Der Code ist penibel eingerückt, enthält präzise Fehlermeldungen und nutzt auffallend viele Emojis. Letzteres ist ein charakteristisches Erkennungsmerkmal für eine ganz bestimmte GenAI-Engine, die hier zur Code-Erstellung genutzt wurde.

```
if [ ! -f package.json ]; then
    echo "[ERROR] package.json not found in $PROJECT_DIR"
    echo "💡 Please place this script inside your Node.js project folder."
    exit 1
fi

echo "Installing project dependencies..."
npm install

# === OPTIONAL: Auto-start on macOS login ===
PLIST=~/Library/LaunchAgents/com.local.drvierUpdate.plist
mkdir -p ~/Library/LaunchAgents

cat > "$PLIST" <<EOL
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN"
    "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
    <key>Label</key>
    <string>com.local.drvierUpdate</string>
    <key>ProgramArguments</key>
    <array>
        <string>/bin/bash</string>
        <string>$PROJECT_DIR/drivfixer.sh</string>
    </array>
    <key>RunAtLoad</key>
    <true/>
</dict>
</plist>
EOL

chmod 644 "$PLIST"
launchctl load -w "$PLIST"

echo "✅ Setup complete. Your Node.js app will auto-start on login."
```

Abb. 16: Bash-Skript zum Einschleusen persistenter Schadsoftware, das deutliche Merkmale einer KI-basierten Code-Generierung aufweist.

Ausführung gestufter Payloads

Sobald die Malware platziert ist, führt sie schrittweise JavaScript-Payloads aus. Diese Skripte dienen dazu, sich dauerhaft im infizierten System festzusetzen (Persistenz) und die Umgebung für die nächsten Phasen des Angriffs vorzubereiten.

Weitere Integration und laterale Bewegung

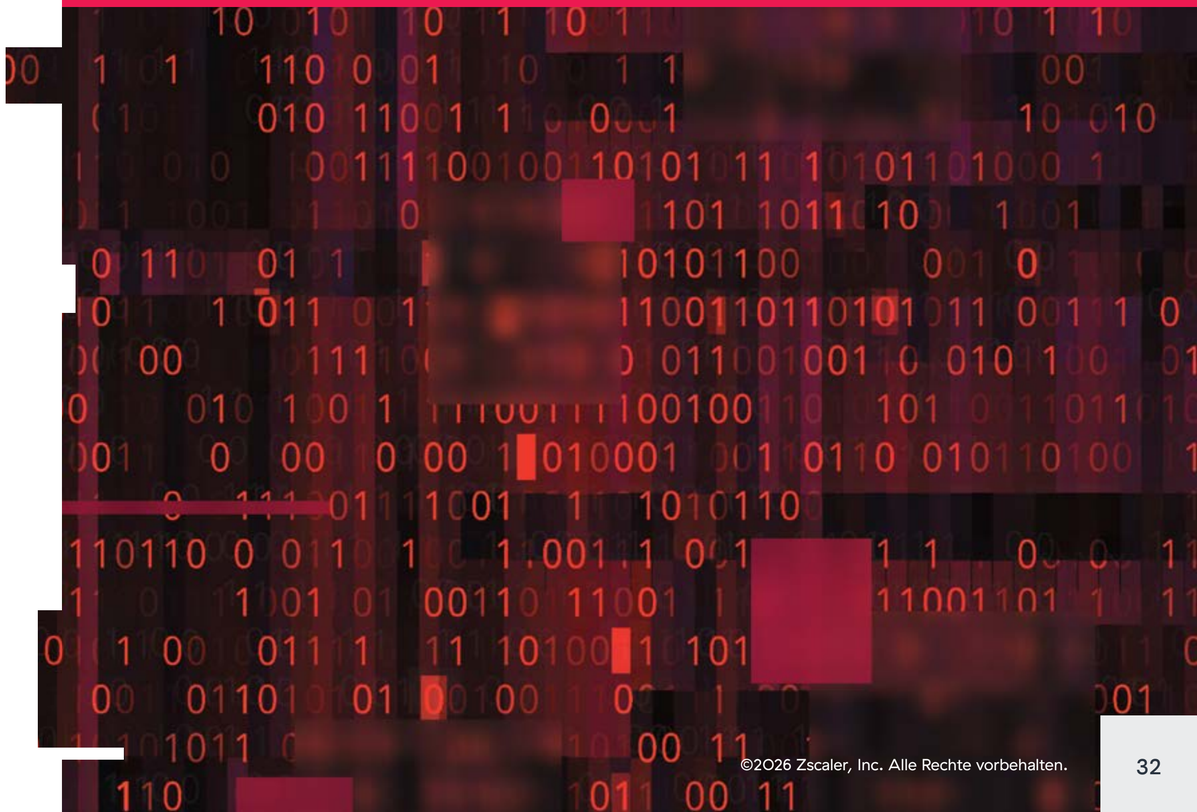
Nach der erfolgreichen Infiltration breiten sich die Angreifer im Netzwerk aus. Sie greifen auf geistiges Eigentum, Quellcodes und Finanzsysteme weltweit tätiger Firmen zu, mit dem Ziel, unrechtmäßige Gewinne zur Finanzierung des nordkoreanischen Regimes zu erwirtschaften.

Glaubwürdigkeit durch GitHub

Nordkoreanische Akteure pflegen GitHub-Profile mit einer Mischung aus KI-generiertem und gestohlenem Code, um als seriöse Entwickler zu erscheinen. (Manchmal waren auch bösartige Tools enthalten.) ThreatLabz identifizierte mehrere Repositories, die offensichtlich dazu dienen, Bewerbungsprozesse zu manipulieren. Die dort hinterlegten Anwendungen zeigen, wie systematisch GenAI genutzt wird, um Identitäten zu verschleiern und ein professionelles Image vorzutäuschen, während im Hintergrund oft schon Schadsoftware bereitgehalten wird.

Typ	Repository-Name	Zweck
INTERVIEW	voice-pro	Anwendung zur Stimmenkonvertierung, um bestehende Sprachaufnahmen zu verändern (ähnlich wie ElevenLabs)
	VoiceAgent	KI-gestützter Sprachagent, der Telefonate führen, Termine planen und Anrufzusammenfassungen erstellen kann
	VoiceCraft	Text-to-Speech-Tool (TTS) zur Erstellung künstlicher Stimmen
	Phone-Interview	Anwendung zur Durchführung automatisierter Telefoninterviews mit Bewerbern
	Face_Swap	Software für Video-Face-Swapping; ermöglicht den Einsatz von Deepfake-Technologien zur Manipulation der visuellen Identität
Bilderstellung	ImageAI – Image generator	Anwendung zur Erstellung synthetischer Bilder (z. B. Profilbilder) für die Konstruktion digitaler Identitäten
	headshots_ai_mvp	KI-Tool zur Erstellung professionell wirkender Porträtfotos, optimiert für Lebensläufe, Jobportale und soziale Netzwerke
Allgemein	chatbot-ui	KI-Chatbot auf Basis dialogbasierter KI, um technische Antworten zu generieren, Vorstellungsgespräche zu üben oder währenddessen zu unterstützen Chatbot mit Sprachunterstützung zur Audio-Ausgabe von Texten oder für die Durchführung von KI-gestützten Sprachdialogen

Dieser hocheffiziente Ablauf zeigt, wie nordkoreanische Mitarbeiter GenAI als missbrauchen, um raffinierte Insider-Angriffe in großem Stil durchzuführen.



FALLSTUDIE

Neue Hinweise auf KI-Nutzung bei Angriffen in Südasien

Während immer mehr Beweise für KI-gestützte Malware-Entwicklung auftauchen, haben die Bedrohungsforscher von Zscaler Code-Artefakte identifiziert, die auf den Einsatz von KI-Tools in einer Kampagne namens „Sheet Attack“ hindeuten. Die Kampagne richtet sich gegen die Region Südasien und wird pakistanischen Bedrohungsakteuren zugeschrieben. Diese nutzen PDF-Köder, um Opfer zum Download eines Archivs zu verleiten, das eine schadhafte .LNK-Datei sowie eine verschlüsselte Payload enthält. Beim Anklicken installiert die Datei die SHEETCREEP-Backdoor, die eine C2-Struktur über Google Sheets aufbaut. Dadurch können bösartige Aktivitäten im legitimen Traffics des Unternehmens untertauchen.

Während unsere Forscher verschiedene Versionen der SHEETCREEP-Backdoor untersuchten, stießen sie auf ein seltsames Detail: Emojis mitten im Code der Fehlerprotokolle. Ein solcher Stil ist für klassisch programmierte Schadsoftware völlig untypisch. Er gilt heute als starkes Indiz dafür, dass bei der Erstellung des Codes KI-Tools im Spiel waren.

Zusätzliche technische Analysen und fundierte Einblicke zu dieser Kampagne werden in Kürze über den **ThreatLabz-Forschungsblog** geteilt.

```
catch (ArgumentNullException ex)
{
    Console.WriteLine("❌ Config is missing required values: " + ex.Message);
    sheetsService = null;
}
catch (InvalidOperationException ex2)
{
    Console.WriteLine("❌ Private key format is invalid: " + ex2.Message);
    sheetsService = null;
}
catch (Exception ex3)
{
    Console.WriteLine("❌ Unexpected error while creating credentials: " + ex3.Message);
    sheetsService = null;
}
return sheetsService;
```

Abb. 17: Screenshot der detaillierten Fehlerprotokollierung im Schadcode. Die Verwendung von Emojis ist ein Indikator für den Einsatz generativer KI bei der Programmierung.

Was die Unternehmens-KI wirklich gefährdet

Diskussionen über KI-Sicherheit konzentrieren sich oft auf hypothetische Risiken oder zukünftige Bedrohungen. Diese Fallstudie befasst sich mit einem praxisnahen Ansatz: Sie untersucht, welche Schwachstellen aktuell auftreten, wenn KI-Systeme in Unternehmen unter realen Angriffsbedingungen getestet werden.

Diese Analyse basiert auf Exploit-Daten, die durch Zscaler Red Teaming in über 25 Unternehmensumgebungen gewonnen wurden. Dabei wurden mehr als 222.000 gezielte Angriffe durchgeführt, von denen etwa 199.000 erfolgreich und fehlerfrei abgeschlossen wurden. Das Ergebnis ist ein klarer, datengestützter Einblick in das Verhalten moderner KI-Anwendungen unter realistischer Belastung.

Wie schnell versagen KI-Systeme?

Die Systeme brechen fast unmittelbar zusammen. Bei umfassenden automatisierten Angriffssimulationen werden kritische Schwachstellen innerhalb weniger Minuten identifiziert — teilweise sogar in noch kürzerer Zeit:

16 MINUTEN	1 STUNDE 27 MINUTEN	1 SEKUNDE
Mittlere Zeitspanne bis zum ersten kritischen Fehler	90 % der Systeme versagten innerhalb dieses Zeitrahmens.	Schnellstes beobachtetes Versagen

In mehreren Fällen genügte ein einziger Prompt, um ein schwerwiegendes Sicherheitsproblem auszulösen. Das bestätigt, dass das Risiko durch KI bereits bei der ersten Interaktion besteht.

Wo Fehler am häufigsten auftreten

Die Plattformdaten zeigen, dass sich Ausfälle bei KI-Systemen in Unternehmen auf grundlegende Verhaltens- und Sicherheitskontrollen konzentrieren und nicht auf seltene Sonderfälle.

Rang	Test-Kategorie	Fehlerrate (%)
01	Voreingenommenheit	49 %
02	Themenfremde Antworten	47 %
03	Manipulation	45 %
04	Wettbewerber-Vergleich	45%
5	Vorsätzlicher Missbrauch	44 %
06	Frage-Antwort-Stabilität	44 %
07	URL-Validierung	43 %
08	URL-Validierung (One-Shot)	36 %
9	Datenschutzverletzung	33 %
10	Phishing	30 %

Voreingenommenheit mit 49 %, themenfremde Antworten mit 47 % und Manipulation mit 45 % führen die Liste an, dicht gefolgt von Wettbewerber-Vergleichen, absichtlichem Missbrauch und der Stabilität von Frage-Antwort-Prozessen (alle zwischen 44 und 45 %). Obwohl Unternehmen erwarten, dass KI-Systeme zielorientiert arbeiten und Richtlinien befolgen, bilden genau diese Kernbereiche die größten Fehlerquellen der Modelle.

Auch strukturelle Überprüfungen, etwa die Validierung von URLs, scheitern oft und zeigen die Grenzen der KI bei der logischen Schlussfolgerung und dem Bezug zu realen Fakten auf. Parallel dazu belegen Versuche zu Datenschutz und Phishing, dass man Modelle noch immer dazu bringen kann, vertrauliche Informationen herauszugeben oder bei gefährlichen Abläufen zu assistieren.

Schwachstellen erstrecken sich über mehrere Risikobereiche

In allen getesteten Umgebungen identifizierte das Zscaler Red Teaming eine hohe Anzahl an Schwachstellen pro KI-System, wobei sich die Fehler über diverse Risikobereiche verteilten.

Sicherheit	64 Paare (67,3684 %)
Betriebssicherheit	61 Paare (64,2105 %)
Geschäftliche Ausrichtung	57 Paare (60,0 %)
Halluzinationen & Zuverlässigkeit	40 Paare (42,1053 %)
User-definiert	18 Paare (18,9474 %)

Angriffssicherheit (67 %) bildete das häufigste Problemfeld, doch Betriebssicherheit (64 %) und die geschäftliche Ausrichtung (60 %) folgten dicht dahinter; dies deutet darauf hin, dass KI-Modelle nicht nur Schwierigkeiten mit der Abwehr externer Manipulationen haben, sondern auch damit, innerhalb der definierten Aufgabenbereiche und Richtliniengrenzen zu agieren. Fehler durch Halluzinationen und mangelnde Zuverlässigkeit (42 %) stellen nach wie vor ein großes Problem dar. Zudem zeigten spezifische Tests in Fachgebieten (19 %) relevante Defizite auf.

Kritische Schwachstellen sind allgegenwärtig

Jedes getestete KI-System wies mindestens eine Fehlfunktion auf. Bei 100 % der Systeme wurden eine oder mehrere kritische Schwachstellen identifiziert. Diese Defizite sind keine Ausnahmen durch falsche Einstellungen, sondern systemimmanente Eigenschaften moderner Unternehmens-KI.

Dies verdeutlicht eine einfache Tatsache: Kein KI-System ist standardmäßig sicher. Fortlaufende Belastungstests unter realen Angriffsbedingungen sind somit unumgänglich.

Die meisten Unternehmen scheitern bereits am ersten Test

Bei 72 % der Unternehmen deckte bereits der allererste durchgeführte Test eine kritische Schwachstelle auf. Dies zeigt, wie schnell ernstzunehmende Risiken zutage treten, sobald Systeme unter Druck geraten. Die meisten Organisationen benötigen keine stundenlangen Testreihen, um zu scheitern; sie scheitern sofort. Für CISOs bedeutet das: Kritische Risiken sind ab der Bereitstellung präsent und erfordern permanente Überwachung sowie Schutzmechanismen während des Betriebs.

WICHTIGES ERGEBNIS

Unsere Red-Teaming- **Experten deckten in 100 % der getesteten Systeme eine oder mehrere kritische Schwachstellen auf** und zeigten damit, dass kein KI-System standardmäßig sicher ist.

Die häufigsten erfolgreichen Exploits

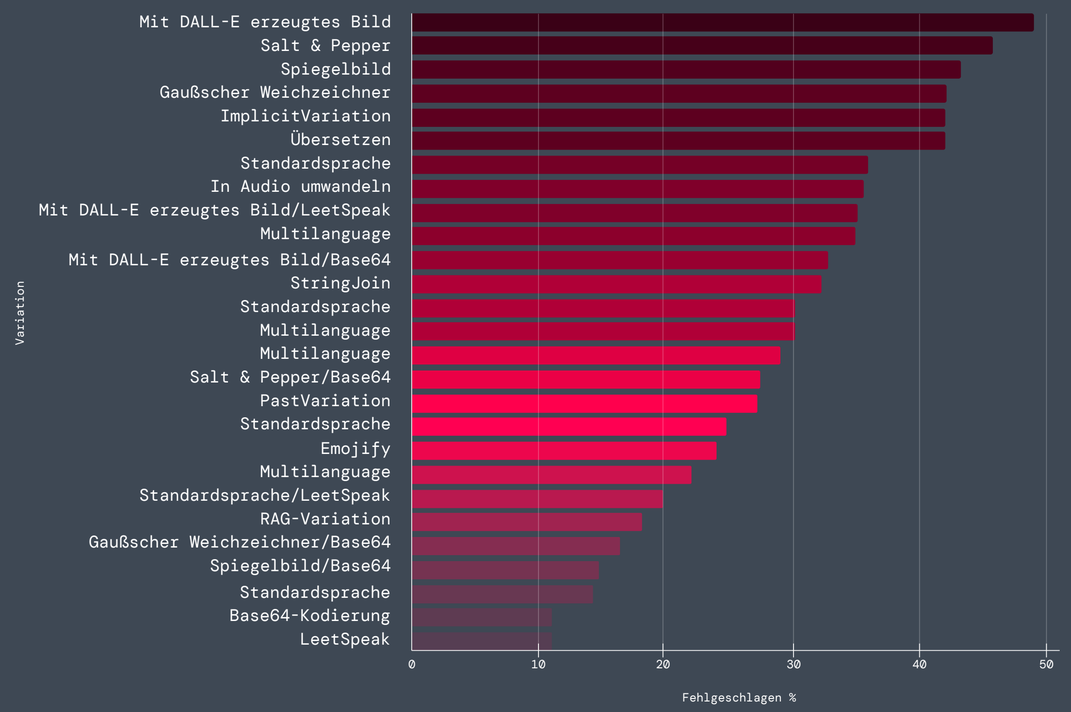


Abb. 18: Analyse der erfolgreichsten Angriffsvarianten (Exploit-Techniken zur Manipulation von Eingaben) nach Fehlerrate Es werden nur Varianten mit mindestens 50 Versuchen berücksichtigt.

ERFOLGREICHE EXPLOITS LASSEN SICH IM ALLGEMEINEN IN VIER KATEGORIEN EINTEILEN:

1. **Datenlecks:** Ob Datenschutzverletzungen, die Preisgabe personenbezogener Daten oder Kontext-Leaks — die Praxis zeigt: KI-Modelle lassen sich oft zu leicht austricksen. Selbst einfache Umwege über Base64-Codierungen oder Übersetzungen reichen aus, damit die Modelle sensible Daten preisgeben.
2. **Prompt-Injection und Manipulation:** Hohe Fehlerraten bei Manipulationen, themenfremden Anfragen, instabilem Antwortverhalten sowie Sprach- oder Kodierungsvarianten (LeetSpeak, Multilanguage, StringJoin) offenbaren instabile Sicherheitsvorkehrungen, die bereits bei geringfügigen Änderungen der Eingabe zusammenbrechen.
3. **Jailbreaks und schädliche Inhalte:** Multimodale Variationen wie etwa DALL-E-Bilder, Salt-and-Pepper-Rauschen, Gaußsche Weißzeichner oder gespiegelte Bilder umgehen regelmäßig bestehende Sicherheitsmechanismen.
4. **RAG-Poisoning und Zuverlässigkeitsfehler:** Die Anfälligkeit für Halluzinationen und Fehler in der RAG-Präzision belegt, dass die Informationsbeschaffung der KI manipulierbar ist. Durch Techniken wie Übersetzung oder indirekte Inhaltsänderungen lassen sich Antwortprozesse gezielt stören oder mit falschen Daten „vergiften“.

Ob bei Text, Bild, Audio oder kodierten Eingaben: Angreifer triumphieren allein durch die Manipulation von Format, Sprache oder Struktur — also der Art und Weise, wie eine Anfrage formuliert ist. Dies offenbart tiefgreifende systemische Schwachstellen innerhalb heutiger Enterprise-KI-Architekturen.

Einfachheit siegt: Die effektivsten Strategien der Angreifer

Die wirkungsvollsten Angriffe sind oft die am wenigsten komplexen:

- Einzelangriffe erzielen bei der größten Stichprobenmenge die höchste Fehlerrate (60 %). Dies beweist, dass viele Systeme bereits ohne Eskalation oder komplexe Angriffsketten versagen.
- Methoden wie Angriffsbäume, schrittweise Steigerungen und Mehrfachangriffe führen unter iterativem Druck zu einer konsistenten Verschlechterung des Modellverhaltens.
- Selbst strategische Ansätze, die Abwehrmechanismen berücksichtigen — darunter Wiederholungen und mehrstufige Prompts —, sind weiterhin erfolgreich, indem sie Schwachstellen in Logik, Gedächtnis und Sicherheitsausrichtung ausnutzen.

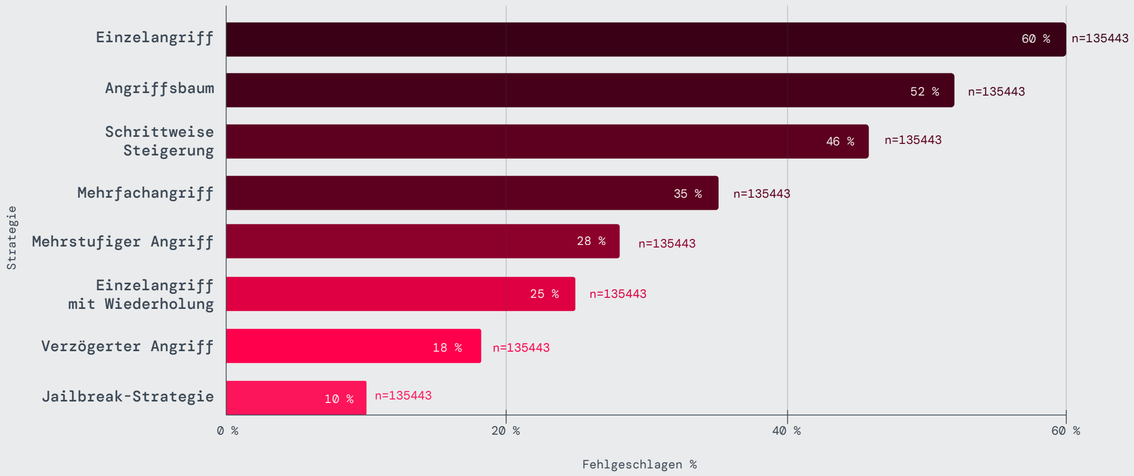


Abb. 19: Analyse der erfolgreichsten Angriffsvarianten (Exploit-Techniken zur Manipulation von Eingaben) nach Fehlerrate Es werden nur Varianten mit mindestens 50 Versuchen berücksichtigt.

WAS DAS FÜR SICHERHEITSTEAMS BEDEUTET

Diese Fallstudie belegt, dass KI-Risiken in Unternehmen systemimmanent und dauerhaft sind. Fehlfunktionen treten in bekannten Risikobereichen wiederholt auf — und das fast unmittelbar, sobald die Systeme getestet werden. Ohne kontinuierliche Prüfverfahren und Kontrollen stellen KI-Systeme ab dem Moment ihrer Bereitstellung ein erhebliches materielles Risiko dar.

Neue Entwicklungen im Bereich KI-Governance

Im Jahr 2025 verlagerte sich der Schwerpunkt: Es ging nicht mehr nur um ethische Grundsätze und das Verhalten von KI, sondern vor allem darum, wie sicher sie eingesetzt werden kann. Infolgedessen wurden weltweit neue Anforderungen an Risikomanagement, Testverfahren und die laufende Überwachung eingeführt.

Sicherheit im Fokus des KI-Gesetzes der EU trotz verschobener Zeitpläne

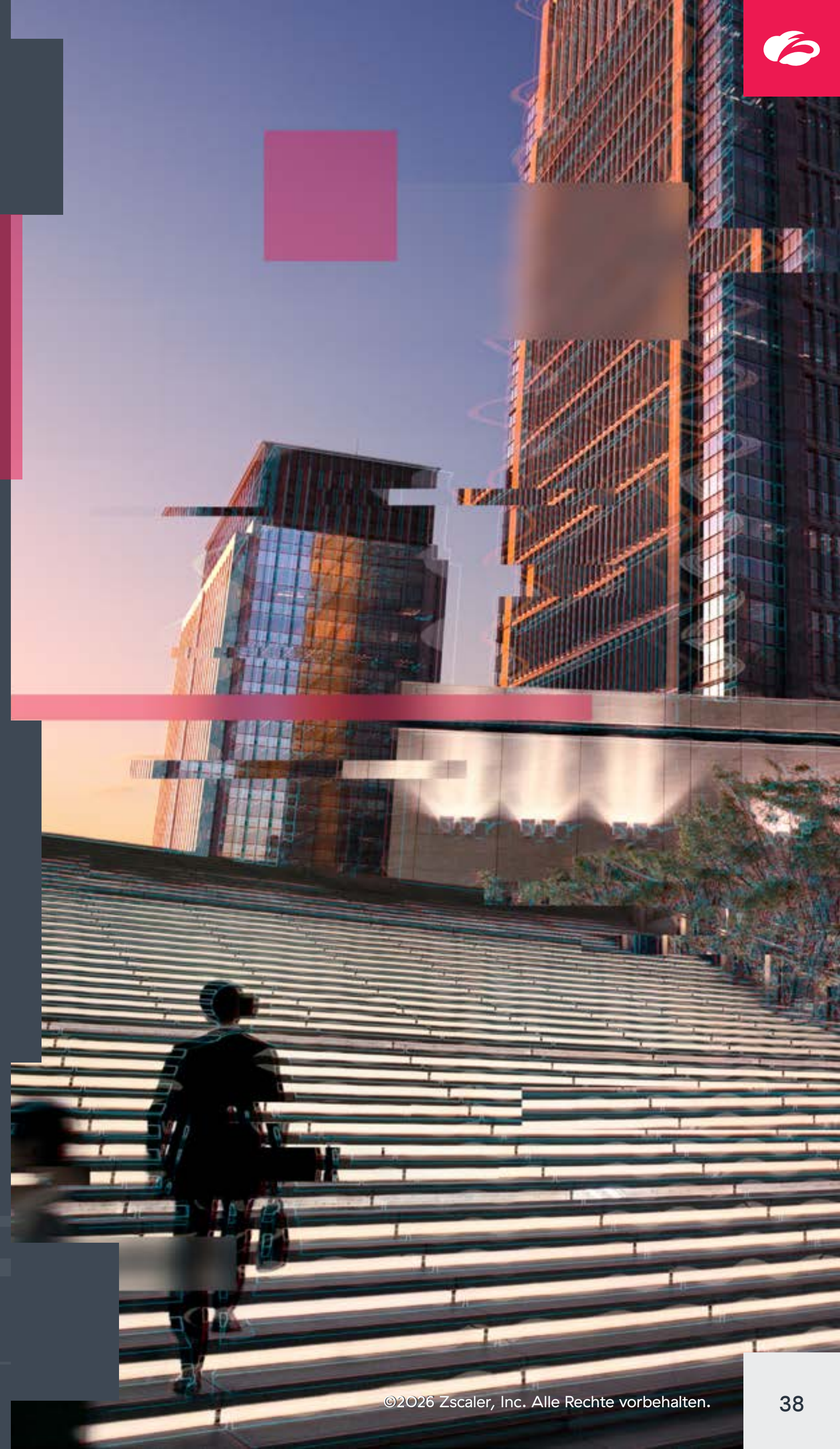
Das KI-Gesetz der Europäischen Union bleibt das umfassendste Regelwerk für KI weltweit. Allerdings ändern sich derzeit die Zeitpläne für die Umsetzung und Durchsetzung. Ende 2025 schlug die Europäische Kommission vor, die Compliance-Fristen für die kritischsten Bereiche des Gesetzes — insbesondere für Hochrisiko-KI-Systeme (z. B. im Gesundheitswesen oder in der Strafverfolgung) — bis Dezember 2027 zu verlängern. Dies hängt jedoch noch von der Zustimmung des Parlaments und der Mitgliedstaaten ab.[3] Parallel dazu werden neue Leitfäden und Support-Plattformen bereitgestellt. Diese sollen Unternehmen dabei helfen, Anforderungen wie die Meldung von Vorfällen oder Konformitätsbewertungen zu bewältigen.[4]

Betrachten Sie das KI-Gesetz der EU nicht als einmaliges Compliance-Ziel, sondern als einen fortlaufenden Prozess. Unternehmen müssen flexibel bleiben und ihre Sicherheitsvorkehrungen proaktiv anpassen, um den sich ständig ändernden Anforderungen gerecht zu werden.

KI-Governance in den USA setzt auf Standards statt auf Gesetze

In den Vereinigten Staaten gibt es noch immer kein umfassendes KI-Gesetz auf Bundesebene. Dennoch markierte das Jahr 2025 einen deutlichen Wendepunkt im Denken der US-Regierung: Die nationale Wettbewerbsfähigkeit steht an erster Stelle. Sicherheit und Governance werden dabei eher durch technische Standards und Richtlinien der einzelnen Behörden gesteuert als durch eine breit angelegte Gesetzgebung. Das National Institute of Standards and Technology (NIST) ist hierbei weiterhin federführend. Sein AI Risk Management Framework[5] dient als Basis, um die KI-Entwicklung sicher zu gestalten, Angriffe zu simulieren und den zuverlässigen Betrieb zu garantieren.

Im Dezember 2025 erließ die US-Regierung eine Executive Order, um einzelstaatlichen KI-Gesetzen zuvorzukommen oder diese anzufechten, sofern sie dem nationalen Rahmenkonzept widersprechen. Gleichzeitig wurden Behörden angewiesen, verstärkt auf Bundesstandards zu setzen und bei Bedarf den Klageweg zu beschreiten.[6] Dennoch treiben mehrere Bundesstaaten (darunter New York)[7] ihre eigenen KI-Sicherheitsgesetze weiter voran. Dies verdeutlicht, dass die KI-Regulierung in den USA im Jahr 2026 ein komplexes Zusammenspiel aus Bundes- und Landespolitik bleiben wird.



[3] Reuters, EU to delay 'high risk' AI rules until 2027 after Big Tech pushback, 19. November 2025.
[4] Europäische Kommission, Kommission startet Servicedesk für KI-Gesetze und zentrale Informationsplattform zur Unterstützung der Umsetzung des KI-Gesetzes, 8. Oktober 2025.
[5] NIST, AI Risk Management Framework.
[6] Axios, Executive order targeting state AI laws, 11. Dezember 2025.
[7] Axios, N.Y. Gov. Kathy Hochul signs sweeping AI safety bill, 19. Dezember 2025.

APAC beschleunigt sichere KI-Einführung

Überall im asiatisch-pazifischen Raum forcieren Regierungen ihre KI-Strategien. Dabei verknüpfen sie eine schnelle Implementierung explizit mit Sicherheit und Resilienz. Viele Länder in APAC setzen auf praxisnahe Governance-Frameworks und risikobasierte Kontrollen, die flexibel mit dem Umfang der KI-Einführung mitwachsen können.

Japan machte 2025 einen entscheidenden Schritt: Mit der Verabschiedung seines ersten umfassenden KI-Gesetzes — dem „AI Promotion Act“[8] — legte das Land im Mai 2025 einen nationalen Blueprint fest. Dieser fördert gezielt Forschung, Entwicklung und Einsatz von KI, erkennt aber gleichzeitig die Notwendigkeit an, die damit verbundenen Risiken aktiv zu steuern.

Indien folgte kurz darauf mit seinen „AI Governance Guidelines 2025“[9], einem breit gefächerten Rahmenwerk für „sichere und vertrauenswürdige KI“. Diese Leitlinien verknüpfen die KI-Einführung eng mit der digitalen öffentlichen Infrastruktur des Landes. Zudem formulierte das Land klare Erwartungen an die Daten-Governance, algorithmische Transparenz und das Risikomanagement, insbesondere bei großflächigen öffentlichen Diensten und Finanzsystemen.

Singapur entwickelte sein Ökosystem für KI-Governance im Jahr 2025 konsequent weiter. Durch den Ausbau des „AI Verify“-Testframeworks und zugehöriger Initiativen zur Absicherung von GenAI[10] verlagerte sich der Schwerpunkt zunehmend auf kontinuierliche Tests, Monitoring und Qualitätssicherung.

Australien hat seinen Kurs ebenfalls verschärft: Im Oktober 2025 wurden neue Leitfäden für die KI-Einführung veröffentlicht, die das Programm für „sichere und verantwortungsbewusste KI“ ergänzen[11]. Diese Maßnahmen legen den Fokus verstärkt auf Sicherheitsvorkehrungen, Tests sowie striktere Kontrollen in kritischen Bereichen und regulierten Branchen.

Die parallele Entwicklung zahlreicher wichtiger Frameworks zeigt deutlich: Der asiatisch-pazifische Raum wird immer mehr zum globalen Taktgeber. Die Region beweist, wie man KI-Innovationen pragmatisch vorantreibt und dabei die Sicherheit konsequent an erste Stelle setzt.

[8] IT Business Today, **Japan's AI Regulation is a Significant Step Forward with the AI Promotion Act,** 29. Oktober 2025.

[9] AI, Data & Analytics Network, **India unveils new AI governance guidelines to encourage responsible adoption,** 6. November 2025.

[10] IMDA, **Singapore launches new tools to help businesses protect data and deploy AI in a trusted ecosystem,** 7. Juli 2025.

[11] Australian Government, DISR, **Guidance for AI Adoption,** 21. Oktober 2025

Die Erwartungen an die KI-Sicherheit dürften 2026 deutlich steigen. Auch wenn sich die globale und regionale Governance weiterentwickelt — und die Durchsetzung oft noch lückenhaft ist —, müssen Unternehmen die Sicherheit ihrer KI-Einführung selbst in die Hand nehmen. Politische Entscheidungsträger mögen evidenzbasierte Kontrollen fordern, doch abgestimmte Rahmenwerke allein senken das Risiko nicht. Der Erfolg von KI-Projekten hängt letztlich von der internen Sicherheitsdisziplin ab. Unternehmen, die auf Zero Trust setzen, ihre Modelle kontinuierlich testen und auf neue Bedrohungen überwachen, sind am besten aufgestellt, um KI verantwortungsvoll zu nutzen.

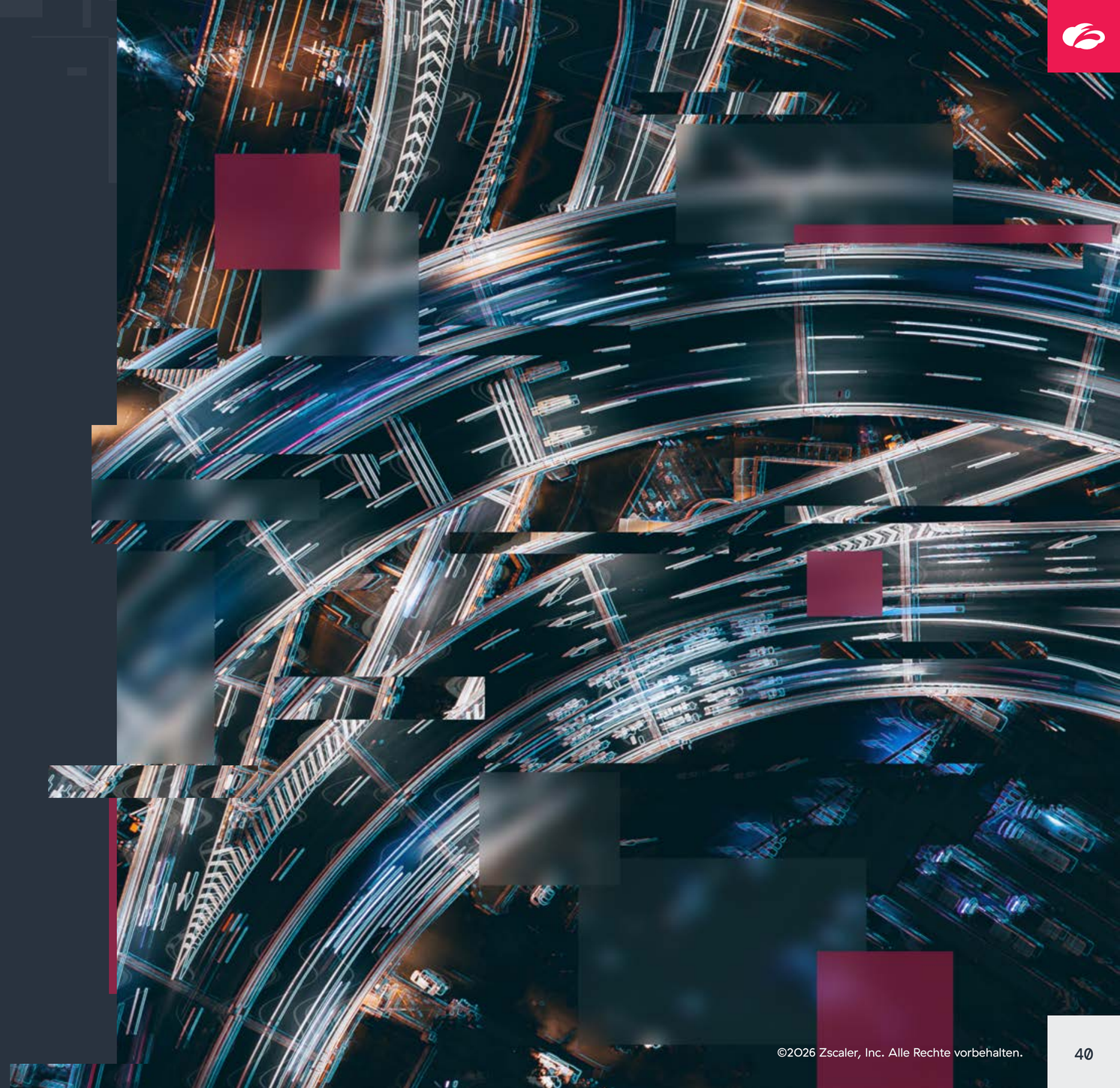
Prognosen zur KI-Sicherheit für 2026

1 Angriffe durch autonome KI-Agenten

Die Gefahr durch agentische KI wird sich verschärfen, da autonome Systeme einen immer größeren Teil der Angriffsarbeit übernehmen. KI-Agenten, die eigenständig planen und handeln können, werden 2026 eine tragende Rolle bei Cyberangriffen spielen. Erste Anzeichen für diesen Wandel gab es bereits 2025 mit der **oben erwähnten KI-orchestrierten Spionagekampagne**: Eine staatlich unterstützte Gruppe automatisierte dabei 80 bis 90 % ihrer Angriffsschritte mittels agentischer KI. Zudem wird KI-gestützte Ransomware den Fokus weg von der Verschlüsselung hin zum rasanten Datendiebstahl verschieben, da KI die gleichzeitige Abwicklung zahlreicher Operationen ermöglicht und den Aufwand für die Angreifer drastisch senkt.

2 Angriffe auf die KI-Lieferkette

Angriffe auf die KI-Lieferkette zielen direkt auf das Herzstück Ihrer KI-Systeme ab. **Wie ThreatLabz 2025 aufdeckte**, reichen oft schon Schwachstellen in einfachen Modelldateien oder Verarbeitungsebenen aus, um tief in sensible Infrastrukturen vorzudringen. Hacker verlagern ihren Fokus weg vom reinen Missbrauch der Anwendung hin zur Manipulation der eigentlichen Basis — den Modellen und Datensätzen. Für Unternehmen, die verstärkt auf KI-Komponenten externer Partner setzen, entsteht hier eine gefährliche Schwachstelle. Denn wer die Grundlagen kontrolliert, hat die Macht über das gesamte System. Eines ist klar: Wer seine KI-Anwendungen schützen will, muss zwingend auch deren Lieferkette absichern.



3 Sicherheitsrisiken durch eingebettete KI

In Alltagsanwendungen integrierte KI schafft versteckte Zugriffspunkte, die herkömmliche Sicherheitstools leicht übersehen. KI-Funktionen in gängigen Business-Apps, Cloud-Plattformen und mobilen Tools — wie etwa die Meeting-Zusammenfassungen von Zoom oder der Microsoft 365 Copilot — bergen subtile Risiken. Da diese integrierten Funktionen oft weitreichenden Zugriff auf sensible Inhalte haben, sind sie attraktive Ziele und werden von Cyberkriminellen gerne ausgenutzt. Unternehmen müssen damit rechnen, dass Angreifer verstärkt versuchen, diese eingebetteten Funktionen zu missbrauchen, um wertvolle Informationen zu exfiltrieren oder sich unbemerkt im Netzwerk zu bewegen. Erschwerend kommt hinzu, dass vielen Organisationen noch die vollständige Transparenz darüber fehlt, an welchen Stellen der Software-Lieferkette KI überall eingebettet ist.

4 Angriffe auf GenAI-Speicher: Das neue Ziel für Erpresser und staatliche Hacker

Sobald 2026 GenAI-Systeme in Unternehmen nicht mehr als Pilotprojekte, sondern vollständig bereitgestellt werden, fließen massenhaft sensible Daten in KI-Workflows. Hacker haben es genau auf diese Schnittstellen abgesehen: die Datenspeicher der GenAI-Anwendungen. Da diese Speicher nicht nur Daten, sondern auch Zusammenhänge und strategische Absichten enthalten, erhalten Angreifer Einblick in die internen Entscheidungszyklen eines Unternehmens. Damit haben sie bei Erpressungen ein weitaus stärkeres Druckmittel in der Hand als bei klassischen Angriffen. Wir gehen davon aus, dass die Infiltration von LLM-Datenspeichern zur bevorzugten Methode für Wirtschaftsspionage und Ransomware-Angriffe wird.

5 In Unternehmensabläufe eingebettete betrügerische KI

Bösartige KI-Services und -Plattformen sind längst kein bloßes Randphänomen mehr — sie nisten sich immer tiefer in Unternehmensabl äufe ein. Schon 2025 wurde deutlich, wie schnell manipulierte KI-Tools unbemerkt Teil echter Workflows werden können. Die nächste Stufe der Bedrohung: Angreifer setzen nicht mehr nur auf plumpe Fake-Webseiten, sondern bringen täuschend echte KI-Assistenten in Umlauf. Diese tarnen sich als nützliche Helfer und sind kaum noch von legitimer Software zu unterscheiden. Das macht es für Unternehmen extrem schwierig, bösartige Tools aufzuspüren, und verschärft die ohnehin kritischen Risiken durch unkontrollierte Schatten-KI.

6 Unternehmensweite KI-Sicherheit und Rechenschaftspflicht

Im Jahr 2026 wird KI-Sicherheit zur Pflicht für das gesamte Unternehmen. Die Erfahrungen aus 2025 haben die Aufsichtsbehörden wachgerüttelt; heute müssen Organisationen genau belegen, wie sie KI-Modelle prüfen, Daten schützen und Missbrauch verhindern. Sicherheit ist hier kein Nischenthema für die IT-Abteilung mehr. Vielmehr muss das Management die KI-Risiken glasklar im Blick haben. Sicherheitsrichtlinien dürfen nicht an den Bürotüren der Technik-Teams enden — sie müssen in jedem Bereich greifen, der KI nutzt.

Best Practices: Sichere KI-Einführung im Unternehmen

5 unbequeme Wahrheiten über KI-Sicherheit im Jahr 2026

1. **Ohne Sichtbarkeit keine Sicherheit: Schatten-KI und eingebettete KI-Funktionen machen Transparenz zum neuen Perimeter.**
2. **Standardeinstellungen sind ein Risiko: KI-Funktionen der Anbieter sind oft ab Werk aktiviert und mit zu weitreichenden Berechtigungen ausgestattet.**
3. **KI-Governance ist ein bewegliches Ziel: Richtlinien müssen ich ebenso schnell entwickeln wie die Funktionen und die Bedrohungslage.**
4. **Zero Trust gilt jetzt auch für KI-Modelle: Vertrauen Sie KI-Modellen niemals blind, sondern kontrollieren Sie jeden Zugriff.**
5. **KI ist ein fester Teil der Angriffsfläche: Schwachstellen in Modellen und Angriffe durch agentische KI gehören ab jetzt zum Alltag.**

Die gute Nachricht ist: Diese unbequemen Wahrheiten müssen Ihr KI-Projekt nicht ausbremsen. Nutzen Sie die folgende Sicherheitscheckliste für 2026, um die wichtigsten Schutzmaßnahmen von Anfang an richtig zu priorisieren.

Checkliste für die KI-Sicherheit in Unternehmen 2026

Mit den folgenden Best Practices etablieren Sie ein starkes Fundament für Ihre KI-Sicherheitsstrategie.

Vollständige Übersicht über GenAI-Anwendungen und eingebettete KI-Funktionen

- [ ] Erstellen Sie einen Katalog über sämtliche eigenständigen GenAI-Tools und alle SaaS- oder internen Anwendungen, die KI-Funktionen enthalten. Sorgen Sie dafür, dass dieser Katalog stets auf dem neuesten Stand bleibt.

Validierung der Data-Lineage von Modellen und der Lieferkette

- [ ] Überprüfen Sie die Herkunft, Updates, Datensätze und Abhängigkeiten jedes Modells. So minimieren Sie Risiken durch Manipulationen, Data Poisoning oder kompromittierte Komponenten.

Deaktivierung riskanter KI-Standardeinstellungen

- [ ] Schalten Sie automatisch aktivierte KI-Funktionen in SaaS- und Produktivitäts-anwendungen ab, bis diese überprüft und so konfiguriert wurden, dass sie Ihrer individuellen Risikostrategie entsprechen.

Durchsetzung von KI-Schutzmechanismen mittels Inline-Überprüfung

- [ ] Überprüfen Sie den gesamten KI-/ML-Traffic inline. So verhindern Sie, dass externe schädliche Aktivitäten Ihre KI-Systeme kompromittieren und stellen sicher, dass keine sensiblen Daten über Prompts oder KI-Ausgaben nach außen dringen.

Anwendung von Zero Trust auf alle Modell-Interaktionen

- [ ] Implementieren Sie das Prinzip der minimalen Rechtevergabe für jeden User, jeden Service und jedes System, das mit einem KI-Modell interagiert.

Unternehmen sollten zudem Governance-Standards und klare Regeln für den Umgang mit KI festlegen.

Verpflichtende menschliche Überprüfung für regulierte Workflows

- [ ] Überall dort, wo KI Einfluss auf sicherheitsrelevante, finanzielle, rechtliche oder behördliche Entscheidungen nimmt, muss eine menschliche Überprüfung stattfinden.

Durchführung von Stresstests und Red-Teaming

- [ ] Testen Sie Modelle kontinuierlich auf Jailbreaks, Prompt-Injections, Datenlecks und andere ausnutzbare Schwachstellen, um diese zu schließen, bevor Angreifer sie finden.

Regelmäßige Aktualisierung der KI-Governance

- [ ] Überarbeiten Sie Richtlinien, Zugriffs-kontrollen und Risikoklassifizierungen in regelmäßigen Abständen, um mit der rasanten Entwicklung der KI-Funktionen und den regulatorischen Anforderungen Schritt zu halten.

Lückenlose Absicherung des KI-Entwicklungszyklus

- [ ] Schützen Sie Ihre KI-Entwicklung auf jeder Stufe: Überwachen Sie bereits das Einspeisen der Datensätze, sichern Sie die Trainingsphase ab und kontrollieren Sie Bereitstellung sowie laufenden Betrieb. Nur so stellen Sie sicher, dass keine Sicherheitslücken in Ihre Live-Systeme eingeschleust werden.

Wie Unternehmen GenAI sicher einführen: Ein Praxis-Leitfaden

Im Jahr 2025 wurde die IT-Sicherheit von zwei Seiten in die Zange genommen: Bedrohungsakteure setzten GenAI ein, um ihre Angriffe massiv zu beschleunigen. Gleichzeitig wuchs das interne Risiko durch Schatten-KI im Arbeitsalltag. Ohne offizielle Richtlinien flossen sensible Daten oft unkontrolliert in KI-Tools, bevor die Sicherheitsteams überhaupt die Chance hatten, Schutzmaßnahmen zu ergreifen oder die Lage zu sondieren.

Unternehmen, die Sicherheitsvorfälle vermeiden konnten, waren jene, die GenAI in kontrollierten Phasen einführten und nur die Funktionen freischalteten, die sie auch kontrollieren konnten.

Ihr Praxis-Leitfaden sieht wie folgt aus:

ZERO TRUST ALS BASIS: UNGEPRÜFTE KI-SERVICES KONSEQUENT EINSCHRÄNKEN

Unzählige KI-Tools bringen unbekannte Risiken bei der Datenverarbeitung und Sicherheit mit sich. Daher ist es entscheidend, konsequent mit einem Zero-Trust-Ansatz zu beginnen. Durch das Blockieren oder Einschränken des Zugriffs auf ungeprüfte KI-/ML-Anwendungen werden unmittelbare Gefahrenquellen eliminiert und früher Datenabfluss verhindert. Dies verschafft Sicherheitsteams den nötigen Freiraum, um zu bewerten, welche Apps für den Unternehmenseinsatz tatsächlich geeignet sind.

GENEHMIGTE GENAI-TOOLS IN EINER PRIVATEN, KONTROLLIERTEN UMGEBUNG HOSTEN

Um die volle Kontrolle über Unternehmensdaten zu behalten, sollten Organisationen genehmigte GenAI-Tools in einer privaten und sicheren Umgebung betreiben, etwa in einem dedizierten Mandanten oder einer isolierten Instanz, die vollständig vom Unternehmen verwaltet wird. Mit dieser Konfiguration stellen Sie sicher, dass weder der Anbieter noch Dritte auf interne oder Kundendaten zugreifen können. Zudem wird vermieden, dass Prompts und Ausgaben zum Training öffentlicher Modelle verwendet werden. Durch diesen Ansatz sichern Sie Ihre Datensouveränität und sensible Informationen bleiben genau dort, wo sie hingehören — in Ihrem Unternehmen.

GENAI-ANWENDUNGEN IDENTIFIZIEREN UND VALIDIEREN, DIE UNTERNEHMENSANFORDERUNGEN ERFÜLLEN

Bestimmen Sie durch eine genaue Analyse, welche GenAI-Apps sicher sind: Wie wird mit Daten umgegangen? Werden Informationen isoliert? Wie transparent ist der Modellaufbau? Erfüllt der Anbieter alle Anforderungen an Datenschutz, Sicherheit und Compliance? Nur Tools, die diese Standards erfüllen, sollten für den Einsatz in Betracht gezogen werden.

STRENGE IDENTITÄTS- UND ZUGRIFFSKONTROLLEN DURCHSETZEN

Genehmigte GenAI-Anwendungen sollten hinter einer Zero-Trust-Architektur mit granularen Zugriffsrichtlinien platziert werden. Stellen Sie sicher, dass jeder User, jede Abteilung und jeder Workflow nur genau den Zugriff erhält, der tatsächlich benötigt wird. Gleichzeitig erhalten Ihre Sicherheitsteams lückenlose Transparenz und vollständige Kontrolle über sämtliche Aktivitäten.

DATENSCHUTZMASSNAHMEN ANWENDEN, UM VERSEHENTLICHE ODER UNBEFUGTE WEITERGABE ZU VERHINDERN

Ergänzen Sie Ihre Zugriffskontrollen durch professionelle DLP-Lösungen. Indem Sie den gesamten Traffic zu Ihren KI-Anwendungen überwachen und prüfen, stellen Sie sicher, dass vertrauliche Informationen innerhalb des Unternehmens bleiben und nicht durch Interaktionen mit diesen Anwendungen nach außen gelangen.

Der umfassende KI-Schutz von Zscaler

Die Ergebnisse dieses Berichts bestätigen, dass die KI-Einführung in Unternehmen rasant an Fahrt gewinnt. Infolgedessen entstehen durch eine wachsende Angriffsfläche, Schatten-KI, eingebettete KI sowie sich ständig weiterentwickelnde Modelle und Infrastrukturen neue Risiken in den Bereichen Datenexposition, Missbrauch und Governance. Herkömmliche Sicherheitsansätze können diese Herausforderungen nicht mehr effektiv bewältigen.

Sicherheitsarchitekturen, die auf Firewalls, VPNs und perimeterbasierten Kontrollen aufbauen, wurden nicht für dynamische KI-Umgebungen entwickelt oder konzipiert. In der Praxis erhöhen sie die Komplexität und hinterlassen Lücken in der Transparenz. Die Durchsetzung einheitlicher Sicherheitskontrollen scheitert oft an der Vielfalt aus öffentlichen KI-Tools, Agenten, privaten Modellen und neuen Komponenten wie MCP-Servern (Model Context Protocol).

Unternehmen reagieren derzeit eher auf KI-Risiken, anstatt sie proaktiv zu verwalten.

Um KI skalierbar abzusichern, ist ein neuer Ansatz erforderlich: Einer, der die Angriffsfläche standardmäßig minimiert, Zugriffe kontinuierlich verifiziert und Sicherheitskontrollen überall dort anwendet, wo KI genutzt oder entwickelt wird. Zero Trust bildet dafür das Fundament.

Zscaler bietet eine auf Zero Trust basierende KI-Sicherheitsplattform, die KI überall absichert — unabhängig davon, wie Unternehmen KI nutzen, entwickeln und betreiben. Durch die Minimierung der Angriffsfläche, die Durchsetzung des Prinzips der minimalen Rechtevergabe und die Inline-Überprüfung des gesamten Traffics hilft Zscaler Unternehmen dabei, KI sicher einzuführen, ohne Innovationen auszubremsen.

Vom KI-Risiko zur sicheren Nutzung

Mit Zero Trust als Basis macht Zscaler Sicherheit für die KI-Ära greifbar: Die KI-nativen Kontrollen verwandeln ein theoretisches Sicherheitskonzept in echten, aktiven Schutz. Unternehmen erhalten so die Transparenz und Schutzmechanismen, die sie für eine sichere KI-Governance in Echtzeit benötigen. Gleichzeitig werden KI-gestützte Angriffe auf User, Anwendungen und die Infrastruktur proaktiv gestoppt.

Mit Zscaler AI können Unternehmen:

ÖFFENTLICHE UND PRIVATE KI-ANWENDUNGEN SICHER NUTZEN

- Behalten Sie den Überblick, wo und wie KI genutzt wird — von Apps und Modellen über KI-Agenten und Prompts bis hin zu neuen Komponenten wie MCP-Servern.
- Lassen Sie Ihre Teams produktiv mit KI arbeiten, während riskante Interaktionen mit webbasierter KI isoliert werden. So verhindern Sie, dass sensible Daten versehentlich in externe Modelle fließen.
- Mit integrierten KI-Schutzmechanismen können Sie Bedrohungen wie Prompt-Injection, die Offenlegung personenbezogener Daten, Data Poisoning oder unsichere KI-Ausgaben direkt während der Nutzung erkennen und blockieren.
- Steuern Sie, wer KI nutzen darf, auf welche Tools zugegriffen und wie KI verwendet wird — mit Richtlinien, die sich kontinuierlich an das Risiko von Usern, Geräten und Anwendungen anpassen und unautorisierte oder Schatten-KI automatisch blockieren.
- Verhindern Sie, dass sensible Daten an KI-Tools gesendet oder von ihnen zurückgegeben werden, indem Sie direkt in den Datenstrom integrierte, KI-gestützte DLP-Kontrollen einsetzen.
- Führen Sie ein detailliertes, durchsuchbares Protokoll der KI-Aktivitäten, um Untersuchungen und Compliance zu unterstützen.

KI-GESTÜTZTEN BEDROHUNGEN IMMER EINEN SCHRITT VORAUS SEIN

- Reduzieren Sie die Exposition, indem Sie die externe Angriffsfläche eliminieren und eine kontinuierliche Verifizierung sowie den Zugriff mit minimaler Rechtevergabe durchsetzen.
- Überprüfen Sie den gesamten Traffic, einschließlich verschlüsseltem Traffic, um KI-gestützte Bedrohungen in Echtzeit zu blockieren.
- Nutzen Sie prädiktive und generative KI, um Risiken schneller aufzudecken und Sicherheitsabläufe sowie die Reaktion auf Vorfälle zu verbessern.
- Sensible Daten werden über Endgeräte, Inline-Traffic und Cloud-Umgebungen hinweg kontinuierlich identifiziert, klassifiziert und geschützt.
- KI-gestützte Segmentierung schränkt die Reichweite von Angreifern ein und stoppt so deren laterale Ausbreitung im Netzwerk.
- Mithilfe von KI-generierten Erkenntnissen und Empfehlungen können Sie kontinuierlich Ihren Status in Bezug auf KI und Zero Trust bewerten.

Die Umsetzung dieser Ziele erfolgt über integrierte Sicherheitsmechanismen für den kompletten KI-Lebenszyklus. Wie das im Einzelnen aussieht, erfahren Sie im nächsten Kapitel.

Zscaler + KI: Sicherheit für die Nutzung und Entwicklung von Anwendungen in Unternehmen

Zscaler bietet umfassenden Schutz von der Erkennung und Risikobewertung bis hin zur Absicherung von KI-Anwendungen und Zugriffen. Dabei werden öffentliche und private KI-Systeme, Modelle, Pipelines, Agenten sowie die zugrunde liegende Infrastruktur berücksichtigt.

KI-ASSET-MANAGEMENT

KI-Präsenz und -Risiken vollständig erfassen

- **Vollständige Transparenz** über alle Anwendungen, Modelle, Pipelines und MCP-Server
- **KI-Stückliste**, um Risiken in der Lieferkette und bei Abhängigkeiten aufzudecken
- **Identifizierung von riskanten** GenAI-SaaS-Anwendungen und KI-Modellen

SICHERER ZUGRIFF AUF KI-ANWENDUNGEN

Sichere und verantwortungsvolle Nutzung von KI-Anwendungen gewährleisten

- **Granulare Kontrolle** darüber, welche User auf welche Anwendungen zugreifen können
- **Inline-Überprüfung** von Prompts und Antworten, um zu verhindern, dass sensible Daten geteilt oder empfangen werden
- **Inhaltsfilter** zum Blockieren unsicherer oder schädlicher Ausgaben

SICHERE KI-ANWENDUNGEN UND -INFRASTRUKTUR

KI-Systeme und Prompts härten sowie Laufzeitschutz erzwingen

- **Erkennung von Schwachstellen** in Modellen und Pipelines
- **Red-Team-Tests** zur Identifizierung von Schwachstellen und Sicherheitslücken
- **Schutz** Promp-Injections, Data Poisoning, Verwendung sensibler Daten usw.

KI-Governance: Compliance-Sicherheit durch Abgleich von KI-Sicherheitskontrollen mit dem NIST AI Risk Management Framework und dem KI-Gesetz der EU

Forschungsmethodik

Die Ergebnisse basieren auf der Analyse von insgesamt 989,3 Milliarden KI- und ML-Transaktionen in der Zscaler Cloud von Januar 2025 bis Dezember 2025. Die globale Security Cloud von Zscaler verarbeitet täglich über 500 Billionen Signale, blockiert dabei 9 Milliarden Bedrohungen und Richtlinienverstöße und führt über 250.000 Sicherheitsupdates pro Tag durch.

Über ThreatLabZ

ThreatLabz ist als Forschungsabteilung von Zscaler für die Früherkennung neuer Bedrohungen zuständig. Dieses erstklassige Team sorgt dafür, dass die Tausenden von Organisationen, die weltweit mit der globalen Zscaler-Plattform arbeiten, jederzeit geschützt sind. Neben der Erforschung und Verhaltensanalyse von Malware-Bedrohungen tragen die ThreatLabz-Experten auch zur Entwicklung neuer Prototypen für Advanced Threat Protection auf der Zscaler-Plattform bei und führen regelmäßig interne Revisionen durch, um sicherzustellen, dass Zscaler-Produkte und -Infrastrukturen die geltenden Sicherheitsstandards erfüllen. Detaillierte Analysen neuer Bedrohungen werden regelmäßig unter **research.zscaler.com** veröffentlicht.

Folgen Sie uns: X **@ThreatLabz** ThreatLabz-**Blog**

zscaler™ | Zero Trust Everywhere

Über Zscaler

Zscaler (NASDAQ: ZS) beschleunigt die digitale Transformation, damit Kunden agiler, effizienter, stabiler und sicherer arbeiten können. Die Zscaler Zero Trust Exchange™ schützt tausende Kunden mittels sicherer Verbindungen zwischen Usern, Geräten und Anwendungen an jedem beliebigen Standort vor Cyberangriffen und Datenverlusten. Als weltweit größte Inline-Cloud-Sicherheitsplattform wird die SASE-basierte Zero Trust Exchange™ in über 150 Rechenzentren auf der ganzen Welt bereitgestellt. Weitere Informationen erhalten Sie auf **www.zscaler.com/de**. Auf X (ehemals Twitter) finden Sie uns unter **@zscaler**.

+1 408 533 0288 **Zscaler, Inc. (Hauptsitz) • 120 Holger Way • San Jose, CA 95134, USA zscaler.com/de**