heise academy

MyCompany GPT

Architektur, Verantwortung und der Weg zu souveräner KI

MyCompany GPT

Architektur, Verantwortung und der Weg zu souveräner KI

Die Geschwindigkeit, mit der künstliche Intelligenz unseren Arbeitsalltag verändert, ist beispiellos. Zwischen Schatten-KI, rasanten Modellzyklen und immer neuen SaaS-Lösungen stehen Unternehmen heute vor einer strategischen Kernfrage: Wie kann KI echten Wert schaffen – sicher, skalierbar und eingebettet in vorhandene Prozesse? Und noch wichtiger: Wie gelingt dieser Schritt, ohne die Organisation zu überfordern? Wir begleiten diese Entwicklungen seit Jahren und sehen: Nicht die Technologie entscheidet über Erfolg, sondern die Art, wie Unternehmen sie einführen, kultivieren und weiterentwickeln.

Das vorliegende Whitepaper zeigt, warum der Aufbau eines eigenen „MyCompany GPT“ mehr ist als ein IT-Projekt – es ist ein Organisationsentwicklungsprozess. Es geht um Architektur und Souveränität, um klare Berechtigungskonzepte, um Data Ownership und um eine Kultur, die Experimente zulässt, aber Qualität verlangt. Wir beleuchten, wie modulare KI-Plattformen entstehen, wo ihre Grenzen liegen, welche Fallstricke RAG-basierte Wissensintegration birgt und warum Pilotprojekte, Governance-Gremien und Feedbackschleifen über Akzeptanz oder Ablehnung entscheiden.

Die Praxisbeispiele und technischen Tiefenanalysen zeigen sehr konkret, wie Unternehmen KI so einsetzen, dass sie produktiv, sicher und nachhaltig wirkt.

Gleichzeitig räumen wir mit einem verbreiteten Missverständnis auf: KI einzuführen bedeutet nicht, ein Tool auszurollen. Es bedeutet, Menschen zu befähigen, mit neuen Möglichkeiten verantwortungsvoll und souverän umzugehen. Wo Technologie skaliert, muss Kompetenz Schritt halten – im Denken, im Entscheiden und im Gestalten.

Genau dafür steht die heise academy: Wir befähigen Fach- und Führungskräfte, KI nicht nur zu nutzen, sondern strategisch zu meistern. Dieses Whitepaper ist ein Beitrag dazu – ein Wegweiser für alle, die KI nicht als Hype begreifen, sondern als Fundament der nächsten betrieblichen Wertschöpfungsepoche.

Anastasia Weiß

Anastasia Weiß
Content-Managerin E-Learning

Lust auf Weiterbildung zum Thema KI?
All unsere Lerninhalte findest du unter **heise-academy.de**

MyCompany GPT in der Praxis: Architektur, Grenzen und echte Learnings

Von Philip Lorenz
Cloud- & DevOps-Engineer & heise academy Experte

Zwischen Hype und Realität

Künstliche Intelligenz ist längst keine Zukunftsmusik mehr, sondern gelebte Realität in deutschen Unternehmen – ob gewollt oder nicht. Eine vielzitierte Microsoft-Studie zeigte bereits vor über einem Jahr, dass rund sieben von zehn Arbeitnehmenden KI-Werkzeuge ohne offizielle Freigabe ihres Arbeitgebenden nutzen. Diese Zahl dürfte in der Zwischenzeit tendenziell weiter gestiegen sein. Diese als „Schatten-KI" bekannte Entwicklung stellt Unternehmen vor eine unausweichliche Entscheidung und konfrontiert sie mit drei grundlegenden Optionen:

1. **Laufen lassen:** Die Augen verschließen und den „Wilden Westen" der unkontrollierten KI-Nutzung zulassen. Ein Weg, der mit der konstanten Gefahr von Datenabflüssen und dem ungewollten Preisgeben wertvoller Geschäftsgeheimnisse gepflastert ist.
2. **Kaufen:** Auf fertige SaaS-Lösungen wie Microsoft Copilot oder Atlassian Rovo setzen. Dies verspricht eine schnelle Lösung, führt aber oft in die Abhängigkeit von einem einzigen Provider, verbunden mit hohen, starren Lizenzkosten und teils undurchsichtigen Abrechnungsmodellen.
3. **Implementieren:** Den strategischen Weg wählen und eine eigene souveräne KI-Plattform aufbauen – ein sogenanntes „MyCompany GPT".

Dieses Whitepaper, basierend auf praktischen Erfahrungen und persönlichen Einschätzungen, widmet sich der dritten Option. Es beleuchtet die Fallstricke, aber auch die erheblichen Mehrwerte eines MyCompany GPTs. Insbesondere für Unternehmen, die Kosten kontrollieren, technologisch modular bleiben und die volle Souveränität über ihre Daten und die eingesetzten KI-Modelle behalten möchten, bietet dieser Weg entscheidende Vorteile. Wir werden zeigen, wie ein solcher Ansatz nicht nur die Risiken der Schatten-KI minimiert, sondern auch maßgeschneiderte Lösungen schafft, die weit über die Funktionalität von Standardprodukten hinausgehen.

Vorteile eines MyCompany GPTs im Vergleich

Entscheidet sich ein Unternehmen für den strategischen Aufbau einer eigenen KI-Plattform, investiert es nicht nur in eine Software, sondern in Souveränität, Kontrolle und langfristige Wirtschaftlichkeit. Im Vergleich zu fertigen SaaS-Produkten ergeben sich entscheidende Vorteile:

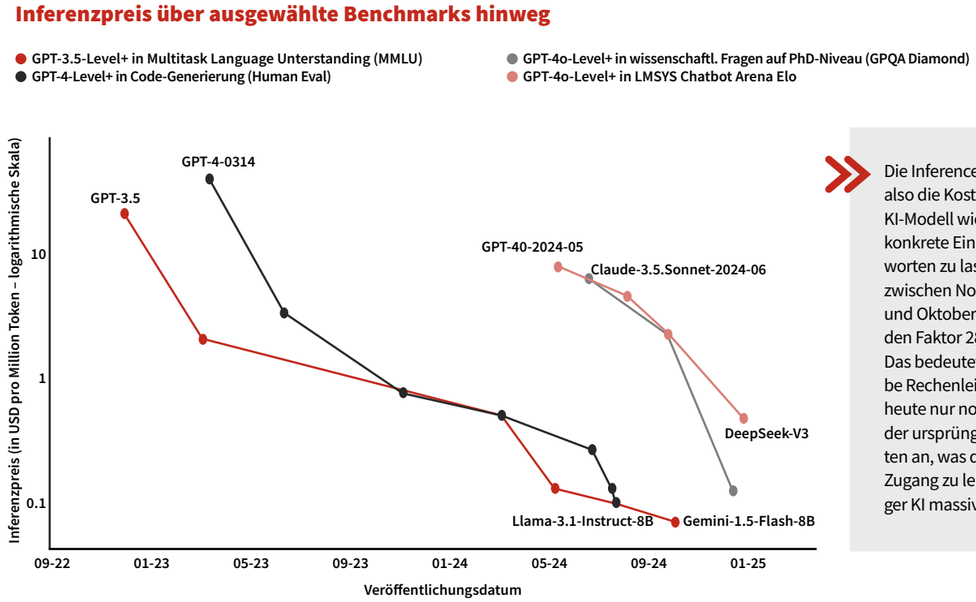
1. Kostenkontrolle und transparente Skalierbarkeit
Während z. B. Microsoft für seinen Copilot-Dienst pauschal rund 30 US-Dollar pro Nutzenden und Monat verlangt, basiert das Kostenmodell eines MyCompany GPTs auf der tatsächlichen Nutzung. Dies ist ein fundamentaler Unterschied. Der KI-Markt entwickelt sich rasant, und die Preise für den Aufruf von Sprachmodellen (LLMs) sinken tendenziell. Bei einer eigenen Lösung schlagen sich diese Preissenkungen direkt in den Betriebskosten nieder.

Im Gegensatz dazu ist es unwahrscheinlich, dass Anbieter von SaaS-Lösungen ihre einmal etablierten Abonnementpreise senken. Stattdessen werden sie versuchen, die hohen Kosten durch zusätzliche Leistungen mit fragwürdigem Mehrwert zu rechtfertigen – selbst wenn sich dies in der Praxis darauf beschränkt, den Nutzenden Credits für die Generierung von Bildern zu geben, die eher für eine humorvolle PowerPoint-Folie als für die tägliche Wertschöpfung relevant sind.

Ein weiteres erhebliches Hindernis bei fertigen Bezahllösungen ist die mangelnde Planbarkeit der Kosten. Anbieter wie Atlassian setzen auf Credit-Systeme, die auf dem Papier einfach aussehen. Die Tücke liegt jedoch im Detail. So gibt Atlassian an, die Kosten pro Aktion basierten auf deren „Komplexität". Für die Anwendenden und Budgetverantwortlichen ist das ein Sprung ins Ungewisse. Mitarbeitende können vor einer Anfrage unmöglich abschätzen, ob diese als „einfach" oder „hochkomplex" eingestuft wird und somit einen oder Dutzende Credits verbraucht. Die Kostenkontrolle wird zum Ratespiel. Ein MyCompany GPT eliminiert diese Unsicherheit. Da die Kosten direkt an messbare Einheiten wie verbrauchte Tokens gekoppelt sind, lässt sich der finanzielle Aufwand exakt prognostizieren und steuern.

2. Granulare Kontrolle und echte Datensicherheit
SaaS-Lösungen verfolgen oft einen All-or-Nothing-Ansatz. Oftmals wird der gesamte Firmentenant vektori-

Inferenzpreis über ausgewählte Benchmarks hinweg



Quelle: Epoch AI, 2025∞ Artificial Analysis, 2025 | Chart: 2025 AI index report

Die Inference-Kosten – also die Kosten, um ein KI-Modell wie GPT-3.5 auf konkrete Eingaben antworten zu lassen – sind zwischen November 2022 und Oktober 2024 um den Faktor 280 gesunken. Das bedeutet: Für dieselbe Rechenleistung fallen heute nur noch 0,36 % der ursprünglichen Kosten an, was den breiten Zugang zu leistungsfähiger KI massiv erleichtert.

siert, um ihn durchsuchbar zu machen. Zwar berufen sich die Provider darauf, dass der resultierende Wissensgraph die individuellen Nutzerberechtigungen berücksichtigt – was technisch oft auch gut gelöst ist.

Doch hier muss sich jedes Unternehmen folgende Frage ehrlich beantworten: Ist unsere bestehende Berechtigungsstruktur im Tenant wirklich so sauber und präzise gepflegt, dass wir es jeder Person erlauben können, die KI nach potenziell sensiblen Dokumenten wie Budgetblättern oder Kostenstelleninformationen zu fragen? Ein MyCompany GPT erzwingt eine bewusste Auseinandersetzung mit Datenquellen und deren Freigaben. Anstatt auf eine fehleranfällige Automatik zu vertrauen, wird der Zugriff auf Daten gezielt und pro Anwendungsfall definiert, was die Sicherheit signifikant erhöht.

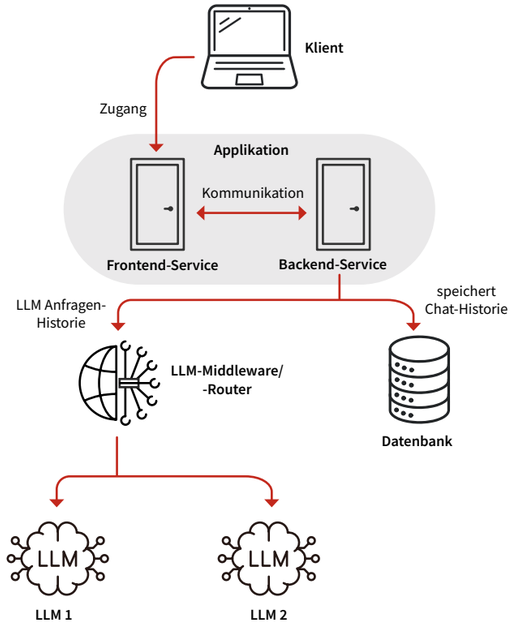
3. Souveränität und technologische Flexibilität

Der größte strategische Vorteil liegt in der Unabhängigkeit. Wer eine SaaS-Lösung kauft, bindet sich an das Ökosystem und die technologischen Entscheidungen eines einzigen Providers und ist dessen Preispolitik, Auswahl an KI-Modellen und Entwicklungs-Roadmap ausgeliefert.

Ein MyCompany GPT hingegen ist modular aufgebaut. Das bedeutet, ein Unternehmen kann jederzeit das für den jeweiligen Anwendungsfall am besten geeignete Sprachmodell auswählen – sei es von OpenAI, Anthropic, Google, Mistral AI oder einem anderen Anbietenden. Diese Flexibilität ist in einem sich schnell verändernden Technologiefeld ein unschätzbarer Vorteil. Ändern sich Lizenzbedingungen, entstehen leistungsstärkere Modelle oder sollen aus Sicherheitsgründen On-Premise-Lösungen genutzt werden, kann die eigene Plattform schnell und ohne Abhängigkeiten angepasst werden. So bleiben Kostenkontrolle und Agilität erhalten.

Architektur: Das Fundament für Flexibilität

Ein entscheidender Vorteil eines MyCompany GPTs liegt in seiner modularen Architektur. Anstatt auf ein monolithisches System zu setzen, wird die Lösung aus entkoppelten, austauschbaren Komponenten aufgebaut. Dieser Ansatz ist keine technische Spielerei, sondern eine strategische Notwendigkeit in einem Umfeld, das sich rasant entwickelt. Neue, leistungsfähigere KI-Modelle, innovative Datenbanktechnologien oder veränderte Lizenzbedingungen erfordern schnelle Anpassungen. Eine modulare Architektur stellt sicher, dass das System agil bleibt und nicht in einer technologischen Sackgasse endet. Die Kernkomponenten und exemplarische Lösungen dafür sind:



- **WebUI mit Chat-Interface:** Benutzeroberfläche, über die Mitarbeitende mit der KI interagieren (z. B. Chainlit, Streamlit, LibreChat etc.).
- **Backend:** Zentrale Anwendungslogik, die Anfragen verarbeitet, Nutzende authentifiziert und die Kommunikation zwischen den Komponenten steuert (oft bereits integriert in Lösungen wie OpenWebUI oder LibreChat, andernfalls Eigenentwicklung).
- **Datenbank:** Speichert Konfigurationen, Chatverläufe und Feedback (MongoDB für Chats/Konfigurationen; Qdrant oder Weaviate für Vektorsuche/RAG).
- **LLM-Middleware:** Abstraktionsebene, die als Weiche zu verschiedenen Sprachmodellen fungiert (z. B. LiteLLM zur API-Vereinheitlichung; LangChain oder LlamaIndex für erweiterte Orchestrierung).
- **LLM:** Eigentliches Sprachmodell als externe API oder selbst gehostete Instanz (z. B. Modelle von OpenAI, Anthropic, Google, Mistral oder Open-Source-Alternativen wie Llama 3).

Diese modulare Anwendungslogik muss auf einem ebenso flexiblen infrastrukturellen Fundament stehen. Hier kommt eine moderne DevOps-Kultur ins Spiel, die konsequent auf Automatisierung und Infrastructure as Code (IaC) setzt. Die Zeiten der „Turnschuh-Administration“, in der Server manuell per ClickOps in einer Weboberfläche konfiguriert wurden, sind vorbei. Auch wenn das initiale Setup mit Werkzeugen wie Terraform oder Bicep zunächst aufwendiger erscheinen mag, ist der Return on Investment (ROI) erstaunlich kurz. Denn im schnelllebigen KI-Umfeld gilt: Die Dinge werden garantiert anders kommen, als ursprünglich geplant. Dank IaC lassen sich Komponenten auf Knopfdruck austauschen, neu aufbauen oder für Testumgebungen replizieren.

Dabei muss es nicht immer gleich der voll ausgebaute Kubernetes-Cluster (K8s) sein. Oftmals lässt sich mit schlankeren Lösungen wie Serverless-Funktionen oder einfachen Container-Services (z. B. Azure Container Apps, AWS App Runner) eine deutlich schnellere „Time to Value“ erreichen. Optimierungen sind kontinuierlich möglich. Die Anwendenden am Ende interessiert es nicht, ob das Chat-Interface auf einem hochkomplexen K8s-Cluster oder einer simplen Container App läuft – Hauptsache, es funktioniert schnell und zuverlässig.

Nutzende und Daten: Der entscheidende Faktor für den Erfolg

Die beste technische Architektur ist wertlos ohne Anwendende, die sie nutzen, und ohne Daten, die ihr einen echten Mehrwert verleihen. Sobald ein erster Prototyp existiert, wächst der Druck, diesen schnell für alle verfügbar zu machen. Doch genau hier ist ein strukturiertes Vorgehen entscheidend, um die Akzeptanz zu sichern und die Kosten im Griff zu behalten.

Pilot: Kontrolliert starten, gezielt lernen

Ein MyCompany GPT sollte sukzessive wachsen – nicht nur hinsichtlich seiner Infrastruktur, sondern vor allem bei den Kosten. Der größte unkalkulierbare Faktor zu Beginn ist der Token-Spent. KI-Kosten entstehen durch die Verarbeitung von Tokens (Input und Output), wobei ein Token in etwa einem Wort oder einer Silbe entspricht. Wie oft und wie umfangreich werden die Mitarbeitenden das System nutzen? Zu Beginn sind das reine Schätzungen.

Ein kontrollierter Pilot mit echten Nutzenden ist daher kein Umweg, sondern eine essenzielle Phase der Datenerhebung. Mit den Nutzungsdaten einer kleinen Gruppe lässt sich ein erster Mittelwert für den Token-Verbrauch bilden, der als Grundlage für realistischere Annahmen und die weitere Skalierung dient.

KI-Gremium als strategisches Zentrum

Um diesen Prozess zu steuern, ist die Gründung eines interdisziplinären Gremiums unerlässlich. Dieses sollte Teilnehmende aus allen relevanten Unternehmensbereichen umfassen. Und ja, auch der Datenschutz muss hier einen prominenten Platz erhalten. Ein guter Rat: Mache die Kolleginnen und Kollegen aus dem Daten- und Compliance-Team frühzeitig zu deinen Verbündeten, denn KI ohne Daten ist wertlos.

Dieses Gremium hat drei zentrale Aufgaben:

1. Auswahl der Nutzenden für die Pilotphase aus verschiedenen Abteilungen.
2. Priorisierung von Anwendungsfällen, die den größten Nutzen versprechen.
3. Identifizierung und Bereitstellung von Datenquellen für diese Anwendungsfälle.

Daten: Treibstoff und größte Hürde zugleich

Mit der Anbindung von Unternehmensdaten ergeben sich die größten Herausforderungen, die über Erfolg oder Misserfolg des gesamten Projekts entscheiden:

1. **Korrektheit der Daten:** Liefert das System auf Basis interner Dokumente falsche oder veraltete Informationen, sinkt die Akzeptanz der Mitarbeitenden extrem schnell. Das Vertrauen ist nur schwer wiederherzustellen.
2. **Zugriffsberechtigungen:** Es muss technisch und organisatorisch sichergestellt sein, dass Mitarbeitende ausschließlich auf Daten zugreifen können, die sie auch sehen dürfen. Ein gängiger Ansatz zur Wissensanreicherung ist RAG (Retrieval-Augmented Generation) über eine Vektordatenbank. Doch hier liegt eine der größten Hürden: Berechtigungskonzepte aus Quellsystemen werden in einer Vektordatenbank nicht immer dynamisch und zuverlässig eins zu eins übernommen.

Der Lösungsansatz hierfür ist organisatorischer Natur: die Benennung von Data Ownern. Für jede angebundene Datenquelle muss es eine klare verantwortliche Person geben. Nur weil es offiziell keine gibt, heißt das nicht, dass sie nicht existiert. Oft müssen diese Personen lediglich gefunden und offiziell ernannt werden. Ohne dieses klare Commitment zur Datenverantwortung sollte keine Datenquelle an das System angebunden werden.

Vom Problem zum Mehrwert: Anwendungsfälle systematisch identifizieren und umsetzen

Sobald ein internes KI-System zur Verfügung steht, führt die Begeisterung oft dazu, dass es als Universallösung für jegliche Probleme gesehen wird. Das KI-Gremium wird schnell mit einer Flut von Anfragen konfrontiert, deren Bandbreite riesig ist: Sie reicht von der Forderung nach komplexen, autonomen Agentensystemen bis hin

zu einfachen Text-Templates – oft im selben Atemzug. Eine der ersten und wichtigsten Aufgaben ist es daher, die Spreu vom Weizen zu trennen und herauszufiltern, welche Probleme wirklich einen Fall für die KI darstellen und welche nicht eher klassische IT-Aufgaben sind.

Strukturierte Interviews für Use-Case-Steckbrief

Um Anwendungsfälle objektiv bewerten und vergleichen zu können, müssen sie in einem strukturierten persönlichen Gespräch erfasst werden. Das Ergebnis ist ein einheitlicher „Steckbrief" für jeden potenziellen Use Case. Dieser sollte mindestens die folgenden Eigenschaften beleuchten:

- **Benötigte Daten:** Welche internen oder externen Datenquellen sind für die Beantwortung der Anfrage notwendig?
- **Data Owner:** Wer ist für diese Daten verantwortlich? Hier gilt die goldene Regel: Ohne eine benannte und engagierte datenverantwortliche Person wird der Anwendungsfall nicht umgesetzt.
- **Bewertung der Daten:** Sind die Daten aus Sicht von Datenschutz und Compliance unkritisch, relevant oder sogar hochkritisch?
- **Volatilität der Daten:** Wie häufig ändern sich die Informationen? (Dies beeinflusst später die technische Architektur, z. B. die Häufigkeit der Indexierung.)
- **Zielgruppe:** Wer genau wird den Anwendungsfall nutzen (z. B. eine Abteilung, eine bestimmte Rolle, das gesamte Unternehmen)?
- **Business Value:** Welchen konkreten Mehrwert liefert der Use Case (z. B. Effizienzsteigerung, Kostenersparnis, Qualitätsverbesserung)? Dieser sollte so messbar wie möglich sein.
- **Technische Voraussetzungen:** Müssen bestehende Systeme oder Schnittstellen angebunden werden?
- **Art der Interaktion:** Handelt es sich um eine reine Datenabfrage, die Ausführung eines Prozesses oder einen komplexen Workflow?
- **Erfolgsmessung:** Wie kann der Erfolg später gemessen werden? Gibt es konkrete KPIs?

Die Kunst der Priorisierung

Während sich die Kosten für Infrastruktur und LLM-API-Aufrufe relativ exakt berechnen lassen, ist der Nutzen oft qualitativer Natur (z. B. Zeitersparnis durch weniger interne Rückfragen). Um den Erfolg des Gesamtprojekts frühzeitig nachweisen zu können, sollten jene Anwendungsfälle bevorzugt werden, deren Nutzen auch klar messbar ist.

Mit der Zeit entsteht so ein genormter Use-Case-Katalog. Dieser dient als Basis für die Priorisierung durch das KI-Gremium.

Dabei haben sich zwei Tipps bewährt:

1. **Use Cases zusammenfassen:** Oft wünschen sich verschiedene Abteilungen ähnliche Funktionen. Diese können gebündelt umgesetzt werden.
2. **Impact maximieren:** Wähle zu Beginn die Anwendungsfälle aus, die den größten positiven Einfluss auf eine möglichst breite Nutzergruppe haben. Das schafft schnell Akzeptanz und verankert das Tool im Unternehmen.

Umsetzung als gemeinsames Unterfangen

Die technische Realisierung eines Anwendungsfalls ist niemals nur eine Aufgabe für die Entwicklerinnen und Entwickler. Es ist ein iterativer Prozess, bei dem alle relevanten Personen an Bord sein müssen:

- Datenschutz und InfoSec zur Freigabe der Daten
- Entwicklerinnen und Entwickler für die technische Umsetzung
- Data Owner als fachlich und inhaltlich Verantwortliche
- Anwendende für Validierung und kontinuierliches Feedback

Gerade der letzte Punkt ist entscheidend. Es muss einen klaren und einfachen Prozess geben, wie die Antwortqualität hochgehalten und stetig verbessert werden kann. Eine Feedback-Funktion direkt in der Chat-Oberfläche (z. B. Daumen hoch/runter, Micro-Feedback) ist hierfür ein essenzielles Werkzeug.

Wissenstransfer: Wie dein Unternehmenswissen in die KI gelangt

Nachdem die Architektur steht und erste Anwendungsfälle definiert sind, stellt sich die zentrale technische Frage: Wie bringen wir dem Sprachmodell bei, was es über unser Unternehmen wissen muss?

Angesichts der enormen Kosten und der tiefgreifenden Expertise, die für das Training eines eigenen Basis-Modells (Foundation Model) von Grund auf erforderlich sind, ist dieser Weg für die allermeisten Unternehmen nicht rentabel. Stattdessen wird auf verfügbare, vortrainierte Modelle etwa von Open AI (GPT-Serie), Google (Gemini-Serie) oder Mistral AI gesetzt.

Die Wahl des Modells und des Hostings ist dabei oft eine pragmatische Entscheidung. Meist wird die bereits vorhandene Cloud-Plattform (Azure, AWS, GCP) bevorzugt, da deren KI-Dienste sofort verfügbar sind und einen schnellen „Time to Value" ermöglichen. Self-Hosting auf lokaler Infrastruktur (On-Premises) ist zwar möglich, geht aber mit hohen initialen Investitionskosten und längeren Vorlaufzeiten einher. Die zuvor beschriebene modulare Architektur stellt jedoch sicher, dass ein späterer Wechsel von der Cloud zu On-Premises – beispielsweise bei verschärften Sicherheitsanforderungen – jederzeit möglich ist.

Unabhängig vom gewählten Modell bleibt eine grundlegende Herausforderung bestehen: Jedes Modell hat einen veralteten Wissensstand. Zusätzlich kennt es weder deine internen Prozesse noch deine aktuellen Produktdaten. Um das Modell nützlich zu machen, muss es an das eigene Unternehmen angepasst werden. Dafür gibt es zwei primäre Methoden: Fine-Tuning und Retrieval-Augmented Generation (RAG).

Fine-Tuning: Modellverhalten formen

Fine-Tuning bedeutet, ein bereits vortrainiertes Sprachmodell gezielt mit zusätzlichen Beispielen nachzutrainieren. Das Ziel ist hierbei weniger die Vermittlung von neuem Faktenwissen, sondern die Anpassung des Modellverhaltens.

Vor- und Nachteile von Fine-Tuning

Vorteile	Nachteile
Domänenspezialisierung: Das Modell lernt Fachjargon, interne Abläufe und die gewünschte Tonalität.	**Hoher Aufwand:** Datenaufbereitung, Qualitätssicherung und Trainingsinfrastruktur sind komplex.
Steuerbares Verhalten: Antworten werden konsistenter und folgen dem gewünschten Stil (z. B. formell).	**Wartungsintensiv:** Das Modell muss regelmäßig neu trainiert werden, um auf dem neuesten Stand zu bleiben.
Offline-Fähigkeit: Kein Live-Zugriff auf externe Daten zur Laufzeit notwendig.	**Intransparent:** Antworten basieren auf internen Modellgewichtungen und sind nicht auf eine Quelle zurückführbar.
Geringere Latenz: Antworten können schneller sein als bei RAG, da keine externe Datenabfrage erfolgt.	**Statisches Wissen:** Das Wissen ist im Modell eingefroren und nicht live an neue Informationen angebunden.
Anpassung für strukturierte Aufgaben: Ideal für spezielle Formate, Tabellen oder Code-Generierung.	**Keine dynamische Kontextanpassung möglich:** Änderungen im Unternehmenswissen erfordern erneutes Training.

Wann Fine-Tuning nutzen?

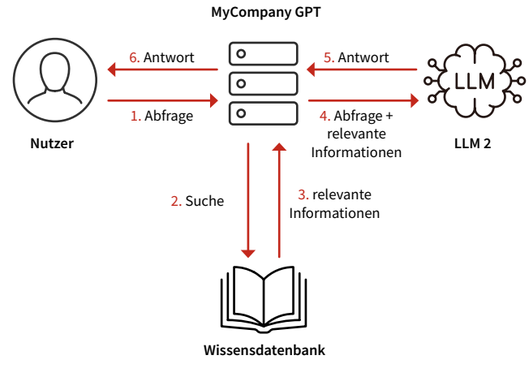
Fine-Tuning ist die richtige Wahl, wenn das Verhalten des Modells angepasst werden soll. Beispiele sind das konsequente Einhalten eines Antwortformats, das Erlernen einer bestimmten Tonalität oder die Spezialisierung auf eine sehr spezifische, repetitive Aufgabe (z. B. das Zusammenfassen juristischer Texte nach einem festen Schema).

Für die Integration von dynamischem Firmenwissen ist Fine-Tuning selten die erste Wahl.

RAG: Dynamische Wissensintegration

RAG ist aktuell der dominierende Ansatz, um LLMs mit externem Wissen anzureichern. Anstatt Wissen fest in das Modell zu trainieren, werden relevante Informationen zur Laufzeit gefunden und als Kontext an die Benutzeranfrage angehängt. Der eigentliche Wissensstand des Modells bleibt dabei unverändert. Das LLM selbst kann dabei jederzeit problemlos und ohne „sunken cost" gewechselt werden.
Der RAG-Prozess lässt sich in drei Kernschritte unterteilen:

1. **Retrieval (Abrufen):** Wird eine Frage gestellt, durchsucht das System eine oder mehrere Datenquellen (z. B. eine Vektordatenbank) nach relevanten Informationen.
2. **Augmentation (Anreichern):** Die gefundenen Informationsschnipsel werden in den Prompt des LLMs injiziert.
3. **Generation (Erzeugen):** Das LLM erhält die Anweisung, die ursprüngliche Nutzerfrage ausschließlich auf Basis des bereitgestellten Kontextes zu beantworten.



Um dies zu veranschaulichen, hier ein konkretes Beispiel, wie eine Anfrage im Hintergrund verarbeitet wird: Ein Beispiel aus der Praxis:

Nutzung der Firmenkreditkarte

1. **Nutzeranfrage (Was der Mitarbeitende eingibt):** Eine Person plant eine Dienstreise und stellt folgende Frage im internen Chat: „Kann ich mit der Firmenkreditkarte das Bahnticket für meine Dienstreise buchen?“
2. **Schritt 1 – Retrieval (Was im Hintergrund passiert):** Das System durchsucht die interne Wissensdatenbank (z. B. eine Vektordatenbank mit Reiserichtlinien und Kreditkartenvorgaben). Es findet folgenden relevanten Auszug: „Auszug aus der Reiserichtlinie – Kapitel 2.3 ‚Zahlungsmittel‘, Version 2.0, Stand 01.04.2025: Für Reisebuchungen wie Bahn-, Flug- und Hotelkosten ist die Firmenkreditkarte zu verwenden. Bei Online-Buchungen über das zentrale Reiseportal wird die Kreditkarte automatisch hinterlegt. Private Auslagen für dienstliche Reisen sind möglichst zu vermeiden.“
3. **Schritt 2 & 3 – Augmentation & Generation (Finaler Prompt an das LLM):** Das System kombiniert nun den gefundenen Kontext mit der ursprünglichen Frage und einer klaren Handlungsanweisung. Der finale Prompt, der an das Sprachmodell (z. B. GPT-4o) gesendet wird, sieht dann so aus:

   --- BEGINN KONTEXT ---
   „Auszug aus der Reiserichtlinie – Kapitel 2.3 ‚Zahlungsmittel‘, Version 2.0, Stand 01.04.2025: Für Reisebuchungen wie Bahn-, Flug- und Hotelkosten ist die Firmenkreditkarte zu verwenden. Bei Online-Buchungen über das zentrale Reiseportal wird die Kreditkarte automatisch hinterlegt. Private Auslagen für dienstliche Reisen sind möglichst zu vermeiden.“
   --- ENDE KONTEXT ---
   Nutzerfrage: „Kann ich mit der Firmenkreditkarte das Bahnticket für meine Dienstreise buchen?“
   Anweisung: Beantworte die Nutzerfrage kurz und präzise. Nutze ausschließlich die Informationen aus dem oben bereitgestellten KONTEXT. Gib die Quelle und deren Stand an.

4. **Ergebnis (Was die Person als Antwort sieht):** Das LLM verarbeitet den angereicherten Prompt und generiert die finale sichtbare Antwort: „Ja, laut Reiserichtlinie (Kapitel 2.3 ‚Zahlungsmittel‘, Stand 01.04.2025) sollen Bahn-, Flug- und Hotelkosten mit der Firmenkreditkarte bezahlt werden. Bei Buchungen über das zentrale Reiseportal ist diese bereits hinterlegt.“

Vor- und Nachteile von RAG

Vorteile	Nachteile
Stets aktuelles Wissen: Informationen sind so aktuell wie die angebundene Datenquelle.	**Performance-Abhängigkeit:** Die Qualität der Antworten hängt stark von der Qualität des Retrievals ab.
Kein Training erforderlich: Es sind keine teuren und aufwändigen Trainingsprozesse für das LLM nötig.	**Keine Verhaltensänderung:** RAG ändert nicht die Tonalität oder das grundlegende Verhalten des Modells.
Transparenz & Nachvollziehbarkeit: Das System kann die Quellen anzeigen, auf denen die Antwort beruht.	**Latenz:** Die vorgeschaltete Suche nach Informationen erhöht die Antwortzeit.
Reduziert Halluzinationen: Das Modell wird angewiesen, bei der Wahrheit (Kontext) zu bleiben, statt zu fantasieren.	**Abhängigkeit von Datenqualität:** Ist die Datenquelle nicht aktuell, ist auch die Antwort nicht aktuell.
Hochgradig skalierbar: Es müssen nur die Kontextdaten gepflegt und erweitert werden, nicht das Modell selbst.	

RAG-Varianten und Abgrenzung zu Function Calling

RAG ist nicht auf die Suche in Vektordatenbanken beschränkt. Gängige Varianten sind:

- **RAG mit strukturierter DB:** Abfragen von SQL-Datenbanken für Produktinformationen oder Kundendaten.
- **RAG mit Web Retrieval:** Live-Abruf von Informationen von Webseiten oder über APIs.
- **RAG + Function Calling (Tool Use):** Dies ist eine besonders mächtige Kombination. Function Calling ist, anders als RAG, ein Mechanismus zur aktiven Interaktion mit externen Systemen (z. B. der Aufruf einer Funktion wie getStockPrice(„AAPL“)). Das Ergebnis dieser Funktion kann dann wiederum als Kontext für RAG verwendet werden.

Einsatzempfehlung für RAG

RAG ist die Methode der Wahl für nahezu alle Anwendungsfälle, die auf dynamischem oder proprietärem Firmenwissen basieren. Es ist ideal für Support-Wikis, Produktdokumentationen, interne Richtlinien und FAQs, da es aktuelle, transparente und nachvollziehbare Antworten ermöglicht.

Der System-Prompt als unsichtbare Leitplanke für verlässliche Antworten

Selbst, wenn das Sprachmodell durch RAG Zugriff auf die richtigen Daten hat, ist damit noch nicht sichergestellt, dass seine Antworten dem gewünschten Stil, Ton und Qualitätsstandard entsprechen. An dieser Stelle kommt eine der mächtigsten aber oft übersehenen Komponenten ins Spiel: der System-Prompt. Er ist die unsichtbare Verfassung des Modells, die ihm vor jeder Interaktion mitgegeben wird und sein Verhalten maßgeblich prägt.

Was ist ein System-Prompt?

Der System-Prompt ist ein spezieller, initialer Satz von Anweisungen, der das Verhalten des LLMs für die gesamte Dauer einer Konversation (Session) steuert. Er wird bei modernen Modellen wie jenen von OpenAI, Anthropic oder Mistral AI vor der eigentlichen Nutzeranfrage ausgeführt. Für die Endanwendenden ist dieser Teil des Prompts nicht sichtbar, doch er ist die entscheidende Leitplanke, die für konsistente und qualitativ hochwertige Ergebnisse sorgt. Er definiert die Rolle, den Stil, die Sprache, die Grenzen und das generelle Aufgabenverständnis des KI-Assistenten.

Die Aufgabe: Vom Hochstapler zum Fachexperten

Sprachmodelle sind von Natur aus höfliche Hochstapler: Sie neigen dazu, lieber eine plausible, aber potenziell falsche Antwort zu erfinden, als zuzugeben, dass sie etwas nicht wissen. Ein gut ausgearbeiteter System-Prompt ist das wichtigste Werkzeug, um diesem Verhalten entgegenzuwirken und das Modell von einem kreativen Generalisten zu einem verlässlichen Fachexperten zu formen.

Seine Hauptaufgaben sind:

- **Konsistenz sicherstellen:** Er sorgt für eine einheitliche Tonalität und Fachlichkeit über alle Antworten hinweg.
- **Rolle zuweisen:** Er weist dem Modell eine klare Rolle zu (z. B. „Du bist ein IT- Support-Mitarbeiter“).
- **Grenzen setzen:** Er begrenzt den Antwortspielraum und verhindert Spekulationen (z. B. „Antworte nur, wenn die Information im bereitgestellten Kontext enthalten ist. Gib andernfalls an, dass du keine Antwort findest.“). Ein System-Prompt trainiert das Modell darauf, im Zweifel zu schweigen, anstatt Märchen zu erzählen.
- **Formatvorgaben machen:** Er kann das Ausgabeformat festlegen (z. B. Markdown, JSON, Tabellen).

Anatomie eines effektiven System-Prompts

Ein guter System-Prompt ist präzise, unmissverständlich und deckt alle relevanten Verhaltensaspekte ab.

Kategorie	Beispielhafte Inhalte
Rolle des Modells	„Du bist ein hilfreicher interner Assistent für die Mitarbeitenden der Mustermann GmbH.“
Ziel/Verhalten	„Beantworte Fragen stets sachlich, präzise und neutral. Deine Hauptaufgabe ist es, Wissen aus den bereitgestellten Quellen zugänglich zu machen.“
Formatvorgaben	„Strukturiere deine Antworten mit Markdown. Nutze Listen und Fettungen zur besseren Lesbarkeit.“
Antwortstil	„Sprich den Nutzenden immer respektvoll mit ‚Sie‘ an.“
Was nicht tun	„Gib keine Ratschläge zu Finanz-, Rechts- oder Medizinthemen. Spekuliere nicht über zukünftige Ereignisse.“
Grenzen setzen	„Antworte ausschließlich auf Basis der Informationen, die dir im Kontext zur Verfügung gestellt werden. Gib an, wenn du Fragen nicht beantworten kannst.“
Sprache/Output	„Antworte ausschließlich auf Deutsch.“
Quellenbezug	„Verweise am Ende deiner Antwort immer auf die zugrundeliegenden Dokumente, die im Kontext genannt wurden.“

Typische Fallstricke: Was vermieden werden sollte

- **Widersprüchliche Regeln:** Anweisungen, die sich gegenseitig ausschließen, verwirren das Modell.
- **Überladene Prompts:** Ein zu langer System-Prompt kann das begrenzte Kontextfenster des Modells unnötig belasten und die Performance beeinträchtigen.
- **Nicht reproduzierbare Logik:** Komplexe, mehrdeutige Anweisungen machen das Verhalten des Modells schwer test- und debugbar.

Fazit: KI braucht Struktur, Strategie – und Geduld

Der Weg zu einem eigenen, unternehmensspezifischen KI-Assistenten ist eine der lohnendsten strategischen Initiativen, die ein Unternehmen heute ergreifen kann. Doch der Erfolg hängt entscheidend davon ab, das Projekt von Anfang an richtig aufzusetzen. Die Reise zum MyCompany GPT endet in vielen Unternehmen in Frustration, wenn sie mit falschen Erwartungen begonnen wird. Ein MyCompany GPT ist kein Plug-and-Play-Produkt.

Auch wenn fertige SaaS-Lösungen wie Microsoft 365 Copilot oder Atlassian Rovo eine schnelle und einfache Implementierung versprechen, entbinden sie Unternehmen nicht von der Pflicht, die eigenen Hausaufgaben zu machen. Ohne eine saubere, gepflegte Daten- und Berechtigungsstruktur wird auch die teuerste Kauflösung unzuverlässige oder sogar gefährliche Ergebnisse liefern. Der Aufbau einer eigenen Lösung erzwingt diese Auseinandersetzung und führt so zu einem nachhaltigeren und sichereren Ergebnis.

Es handelt sich um ein Transformationsprojekt, kein Tool-Deployment. Die Einführung eines KI-Assistenten ist vor allem ein Change-Management-Prozess. Es reicht nicht, eine Technologie bereitzustellen. Mitarbeitende müssen begleitet, Unsicherheiten abgebaut und klare Leitplanken gesetzt werden. Dazu gehört, schnell FAQs und Guides bereitzustellen („Was darf ich, was darf ich nicht?“), Best Practices für gutes Prompting zu vermitteln und eine Kultur der Datenverantwortung (Data Ownership) zu etablieren. Ein zentraler Aspekt ist dabei die Reduzierung der kognitiven Last für die Anwendenden. Wir IT-Fachleute lieben vielleicht Regler und Schalter – der normal Nutzende tut es nicht. Es sollten nicht sechs verschiedene KI-Modelle zur Auswahl gestellt werden; die Komplexität der dahinterliegenden Technologie muss verborgen werden, um sich auf die eigentliche Arbeit konzentrieren zu können.

Wer klug startet, spart langfristig – an Kosten, Frustration und Enttäuschung. Ein überstürzter Start mit unsauberen Daten rächt sich bitter. Sollten Mitarbeitende zu Beginn das Gefühl bekommen, dass die Antworten des Systems oft falsch oder unzuverlässig sind, ist das Vertrauen schnell zerstört und nur mühsam wieder aufzubauen. Ein schrittweises Vorgehen, das auf einer flexiblen Architektur, klar definierten Anwendungsfällen und einem unbedingten Fokus auf Datenqualität und Nutzerfeedback beruht, ist der sicherste Weg zu einem System, das sein volles Potenzial als wertvoller, digitaler Assistent entfalten kann. ≪

7 Tipps für dein MyCompany GPT

1. Gründe ein KI-Gremium.
Steuere das Projekt mit einem interdisziplinären Team aus allen Fachbereichen, inklusive Datenschutz und IT-Sicherheit.

2. Etabliere Data Owner.
Binde keine Datenquelle an, für die nicht eine Person klar die Verantwortung für Qualität und Korrektheit übernimmt.

3. Setze auf modulare Architektur.
Die Fähigkeit, einzelne Komponenten wie das Sprachmodell oder die Datenbank einfach austauschen zu können, ist deine Versicherung gegen die Schnelllebigkeit des KI-Marktes.

4. Nutze RAG statt Fine-Tuning.
Für die Anreicherung von dynamischem Firmenwissen ist Retrieval-Augmented Generation (RAG) fast immer der flexiblere, transparentere und kosteneffizientere Ansatz.

5. Kontrolliere die KI mit einem System-Prompt.
Definiere Rolle, Tonalität und Grenzen des Modells präzise, um zuverlässige und sichere Antworten zu gewährleisten.

6. Verberge die Komplexität.
Biete den Nutzenden eine einfache, intuitive Oberfläche ohne unnötige technische Auswahlmöglichkeiten.

7. Stelle Datenqualität über alles.
Das Vertrauen deiner Mitarbeitenden ist die härteste Währung. Einmal verloren, ist es nur schwer zurückzugewinnen.

Eigenes Unternehmens-GPT aufsetzen: 9 von 10 Unternehmen sollten kaufen statt entwickeln

Von Robin Harbort
Founder, AI-Analyst, KI-Experte & heise academy Experte

Eine große Frage treibt derzeit IT-Teams und KI-Beauftragte um: Sollten wir ein eigenes Unternehmens-GPT bereitstellen, und wenn ja, wie? Ganz klassisch gibt es hier zwei Ansätze: Eine Eigenentwicklung im Unternehmen aufbauen oder eine bestehende generative KI-Lösung einkaufen. In der Fachsprache ist das die klassische Make-or-Buy-Entscheidung. Ich möchte hier klar Stellung beziehen: Setze auf externe KI-Softwarelösungen statt auf Eigenentwicklungen! Warum? Weil sich die Vorteile externer Plattformen – schneller Fortschritt, sinkende relative Kosten und Incentivierung – meiner Erfahrung nach deutlich durchsetzen. Diese Empfehlung gilt vor allem für KMUs und Non-Tech-Firmen. Schauen wir uns die Argumente im Detail an.

Qualität: Rasantes Entwicklungstempo überfordert Eigenbauten

Die Welt der Large Language Models (LLMs) und KI-Anwendungen entwickelt sich in einem atemberaubenden Tempo. Kaum war GPT-3.5 verstanden, wurde es von GPT-4 abgelöst und nun dominiert GPT-5 die Benchmarks. Neue Modelle, Features und Verbesserungen erscheinen wöchentlich. Mit diesem Innovationstempo

können firmeninterne Entwicklungen nicht mithalten. Was heute Start of the Art ist, kann in sechs Wochen veraltet sein. Selbst ambitionierte KI-Teams stoßen hier an Grenzen – es ist frustrierend, wenn ein mühsam intern aufgebautes System plötzlich hinter der Qualität externer Angebote zurückbleibt.
Externe KI-Provider hingegen treiben den Fortschritt unermüdlich voran. Sie bündeln enormes Fachwissen und gigantische Datenmengen. Gepaart mit der Finanzierung durch ausdauernde Wagniskapitelgebend entstehen KI-Applikationen von atemberaubender Qualität. OpenAI, der Macher von ChatGPT, verfügt über unvorstellbare Ressourcen. Gleiches gilt für andere KI-Firmen wie Anthropic, Glean, Harvey oder xAI. Milliarden an Finanzmitteln, Millionen an datenliefernden Kundinnen und Kunden im Ökosystem, Hunderttausende GPUs von NVIDIA und Tausende von Top-Talenten. Bei der internen Umsetzung von KI-Applikationen, wie einem Unternehmens-GPT überschätzen sich Unternehmen gerne.

Jeder Unternehmens-GPT muss denselben Härtetest bestehen: Ist die interne Lösung so gut wie ChatGPT? Die Antwort lautet fast immer: Nein.

Kurz gesagt: Der KI-Markt bewegt sich schneller, als interne Projekte mithalten können. Wer auf eigene Lösungen setzt, riskiert am Ende mit einer veralteten oder qualitativ schwächeren KI dazustehen. Das schwächt die KI-Adoption signifikant. Die externen Plattformen geben dagegen das Tempo vor – und die Unternehmen können und sollten davon profitieren.

Kosten: Deflation von KI-Inferenz meistern nur externe Anbieter

Wer sich mit KI beschäftigt, muss verstehen: KI ist Intelligenz. Und die Kosten für Intelligenz fallen rasant. Externe Anbieter von generativer KI können ihre Infrastruktur, Modelle und Backend-Prozesse kontinuierlich verbessern – und das oft täglich. Sie profitieren von Skaleneffekten, neuen Modell-Architekturen, effizienterer Hardware und besseren Prozessen. Diese Fortschritte geben sie entweder in Form geringerer Preise oder höherer Leistung zum gleichen Preis an ihre Kundinnen und Kunden weiter.

Seit 2022 sind die Kosten für das Verarbeiten von einer Billion Tokens – das entspricht etwa dem 16-fachen Umfang von Wikipedia – von 60 Millionen auf nur noch 60.000 US-Dollar gesunken. Das ist ein Rückgang um 99,9 %. Prognosen zufolge könnten diese Kosten bis 2042 sogar auf unter einen Millionstel US-Dollar fallen.

Externe Anbieter können an dieser Deflation teilnehmen, da sie Modell-APIs wechseln, die neueste und effizienteste Generation von NVIDIA-GPUs einsetzen und State-of-the-Art-Datenquellen haben. Das eigene Unternehmen hat diese Möglichkeiten nur begrenzt. Interne Lösungen hingegen werden oft mit der Zeit relativ teurer, weil sie nicht vom allgemeinen Preisverfall profitieren. Es fehlt an Ressourcen, um laufend mit den besten Standards mitzuhalten. Der notwendige Anpassungsdruck, etwa Modelle regelmäßig neu zu trainieren, Pipelines zu optimieren oder Hardware auszutauschen, ist in der Realität kaum umsetzbar. Was heute noch wirtschaftlich erscheint, kann mit dem nächsten X-Post von Sam Altman überholt und kostenineffizient sein.

Daraus folgt: Eigene KI-Systeme werden mit der Zeit nicht besser, sondern relativ gesehen teurer – einfach, weil sie vom stetigen Preisverfall auf dem Markt entkoppelt sind. Wer hingegen auf etablierte Plattformen setzt, reitet die Welle des technologischen Fortschritts mit. Das macht externe Lösungen nicht nur planbarer und einfacher, sondern auch langfristig wirtschaftlich überlegen.

Incentivierung: Lizenzmodell statt Pay as you go

Schauen wir uns nun das Lizenzmodell an – ein entscheidendes Kriterium für die Incentivierung der Nutzung von KI. Funktioniert eine KI gut, sollte sie maximal eingesetzt werden. Das Lizenzmodell muss hier die Mitarbeitenden richtig incentivieren.

KI-Apps werden häufig im Lizenz- oder Abonnementmodell angeboten. Das heißt, dein Unternehmen zahlt einen festen Betrag (z. B. pro Monat oder pro Nutzenden) und kann die KI dafür innerhalb dieses Rahmens beliebig oft nutzen. Diese Fixkosten sorgen für Planungssicherheit. Je mehr Anwendungsfälle ihr findet und je mehr Kolleginnen und Kollegen die KI nutzen, desto besser verteilt sich dieser Festpreis auf die Nutzung.

Eine Herausforderung , die bei selbst betriebenen Lösungen vorherrscht, entfällt damit: Niemand muss ein schlechtes Gewissen haben, wenn die KI zu viel genutzt wird. Im Gegenteil, ihr habt einen Anreiz, KI breit einzusetzen, denn die Kosten sind bereits gedeckelt. Bei der Eigenlösung sieht das nämlich anders aus. Hier bezahlt ihr meist verbrauchsbasiert – ob ihr nun Cloud-Rechenleistung pro API-Call abrechnet oder lokale Server nach Auslastung betreibt. Jede zusätzliche Nutzung verursacht unmittelbar Mehrkosten. Wird euer KI-Service also populär und die Anfragen schießen in die Höhe, schnellen auch eure Ausgaben hoch. Diese steigenden variablen Kosten wirken wie eine Bremse. Es wird sich zweimal überlegt, ob man die KI wirklich überall einsetzt, weil jede Nutzung das Budget belastet. Ironischerweise kann eine selbstgebaute Pay-as-you-go-KI so den eigenen Erfolg ausbremsen.

Incentivierung ist entscheidend
Wenn die KI gut funktioniert und ich sie jeden Tag mehrfach einsetzen will, möchte niemand von seinem Chef hören: Verwende weniger KI, die ist so teuer.

Make or Buy – die klassischen Kriterien im KI-Kontext

Lass uns die Entscheidung anhand bekannter Kriterien final betrachten:

Kontrolle & Datenschutz
Ja, interne Lösungen geben volle Kontrolle. In streng regulierten Branchen ist das ein Argument. Doch moderne Enterprise-KI-Angebote bieten oft ausreichende Garantien für Datenschutz und Compliance. Vor allem in KMUs, die keine KI-Sicherheits- und EU-AI-Act-Fachleute habe, lassen sich die Sicherheitsanforderungen mithilfe von Marktlösungen um ein Vielfaches besser umsetzen.

Abhängigkeit
Externe Anbieter bedeuten Abhängigkeit. Das ist ein gewichtiges Contra-Argument. Es gibt KI-Anbieter, die mit guten Konditionen locken. Im Rahmen der Abhängigkeit werden die Konditionen mit der Zeit immer schlechter. Gute Vertragsgestaltung und ein Plan B (z. B. Datenexport) reduzieren dieses Risiko beim Buy-Ansatz deutlich.

Fazit: Warum du besser kaufst

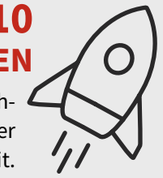
Mein Fazit ist eindeutig: ***Kaufe generative KI-Lösungen ein.*** Du bist schneller, flexibler, wirtschaftlicher – und sicherer unterwegs. Die meisten Unternehmen profitieren so deutlich mehr, als wenn sie mühsam eine eigene Lösung aufbauen.

Natürlich gibt es Ausnahmen – bei enorm sensiblen Daten oder Spezialfällen. Doch für 90 % der Firmen gilt: ***Setze auf bewährte Plattformen*** und nutze deine internen Ressourcen für die Kernprozesse deines Unternehmens, in denen deine Expertise liegt. ≪

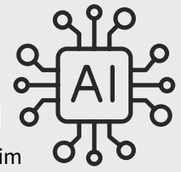
1 VON 3 UNTERNEHMEN
das den Einsatz von KI plant möchte ein LLM selbst betreiben.

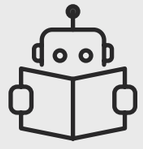
68 % der Unternehmen würden Provider aus Deutschland bevorzugen.

9 VON 10 UNTERNEHMEN
sehen KI als Zukunftstechnologie für den Erhalt der Wettbewerbsfähigkeit.



JEDES 3. UNTERNEHMEN
plant den Einsatz von KI im internen Wissensmanagement.





DIE HÄLFTE
der Unternehmen führt keine Schulungen zu KI durch.

Quelle: Bitkom Research 2025: Digitalisierung der Wirtschaft 2025, Bitkom Research 2024: Künstliche Intelligenz in Deutschland: Perspektiven aus Bevölkerung & Unternehmen

Generative KI im Unternehmen: Sicherheitsrisiken verstehen

Von Frank Ully
Principal Consultant Cybersecurity & Head of CORE bei Corporate Trust Business Risk & Crisis Management GmbH, heise academy Experte

Künstliche Intelligenz, insbesondere in Form großer Sprachmodelle (LLM), wird Wirtschaft, Alltag und Weltpolitik in den kommenden Jahren maßgeblich prägen. Unternehmen müssen diese neue Technologie von Anfang an sicher einsetzen.

Organisationen sollten Fehler der Vergangenheit nicht wiederholen: Sicherheit bei IT, Web, IoT und zuletzt der Cloud wurde oft erst nachträglich, mühsam und lückenhaft angeflanscht. Dies geschah, nachdem die Systeme bereits großflächig und unzureichend gesichert eingeführt waren und sich Sicherheitsvorfall an Sicherheitsvorfall reihte.

Sprachmodelle sind geplagt von bekannten Problemen wie Halluzinationen und Rechenschwächen. Deren Ursache wird verständlich, wenn man die zugrundeliegende Technik versteht: LLMs werden in mehreren Schritten trainiert, zunächst zu Internet-Dokumentengeneratoren und dann zu hilfreichen Assistenten.

Zusätzliche Herausforderungen entstehen, wenn Sprachmodelle mit Werkzeugen ausgestattet werden und so zu Agenten avancieren. Simon Willison, Mitentwickler des Web-Frameworks Django und Entwickler eines eigenen LLM-Kommandozeilentools, spricht von einer „tödlichen Dreierkombination“ von Fähigkeiten: Zugriff auf unternehmensinterne Daten, Kontakt mit nicht vertrauenswürdigen Inhalten und der Fähigkeit, extern zu kommunizieren. Dadurch können Angreifende den Unternehmensagenten dazu bringen, private Daten nach außen zu tragen.

Was können Sicherheitsverantwortliche, Entwicklerinnen und Entwickler und Admins also tun, wenn sie große Sprachmodelle, KI-Automationen oder KI-Agenten einsetzen wollen, dürfen oder müssen – und das auf möglichst sichere Art?

Zunächst müssen Sicherheitsverantwortliche sich bewusst sein, dass selbst bei schnell einsatzbereiten Lösungen wie KI-SaaS, etwa bei den Copiloten von Microsoft, das Modell der gemeinsamen Verantwortung gilt: Für viele Sicherheitsaspekte bleiben die Anwendenden selbst zuständig. Wie bei der Cloud entbindet das Nutzen eines Services nicht von der Verantwortung für die Sicherheit und Sicherung der Unternehmensdaten.

Zweitens müssen Sicherheitsverantwortliche grundsätzlich verstehen, wie diese neuen Technologien funktionieren; was ihre Stärken und Schwächen sind.

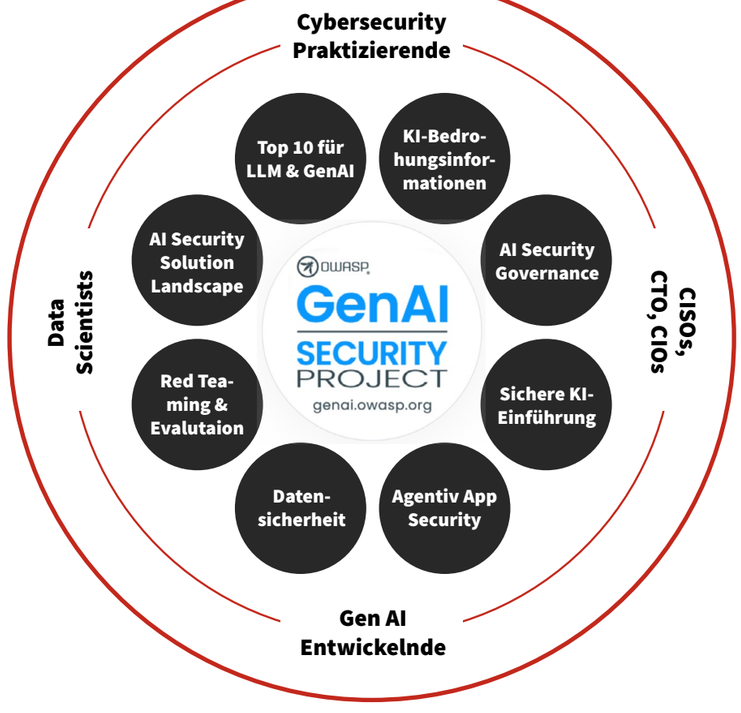
Die CSA veröffentlicht in ihrer Sichere-KI-Initiative zahlreiche kostenlose, lesenswerte Dokumente für das Sicherheitsmanagement, darunter die Reihe „AI Organizational Responsibilities“, Bedrohungstaxonomien und ein Risikomanagement-Framework.

Praktischer wird es bei der OWASP: Das Open Worldwide Application Security Project setzt sich seit 2001 für die Sicherheit von Webanwendungen ein. Seit 2023 kümmert sich die GenAI-Security-Gruppe innerhalb der OWASP um die Absicherung von generativer KI. Auf deren Webseite kann man ihre zahlreichen Veröffentlichungen einsehen, die sich an verschiedene Zielgruppen richten, von CISO/CTO/CIO über Cybersecurity-Fachleute bis hin zu KI-Entwicklern. Hervorzuheben sind drei Dokumente: Die „OWASP Top 10 for LLM Applications“ geben als Awareness-Dokument einen Überblick über die kritischsten Risiken von Prompt Injection über die Preisgabe sensibler Informationen bis hin zu übermäßiger Handlungsfreiheit (hier erkannt man das Willison’sche Dreigespann) und welche Maßnahmen dagegen helfen. Die „LLM AI Cybersecurity & Governance Checklist“ führt in vertrauenswürdige KI ein, beschreibt unterschiedliche Strategien und enthält Prüflisten. Die „LLM and GenAI Data Security Best Practices“ skizzieren auf 60 Seiten bewährte Praktiken.

Daneben haben zahlreiche andere Normungsgremien und Privatunternehmen Rahmenwerke für sichere KI veröffentlicht. Aufgrund von Praxisnähe und Übersichtlichkeit ist Googles Secure AI Framework (SAIF) hier besonders hervorzuheben.

Sicherheitsexperten wie Chris Hughes von Resilient Cyber argumentieren, die Security-Branche könnte diesmal nicht nur rechtzeitig zum Absichern der neuen Technologie zur Stelle sein, sondern Vorreiter beim verantwortungs- und sinnvollen Einsatz von KI werden. ≪

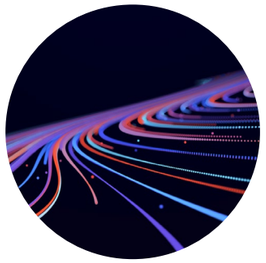
Überblick über das OWASP GenAI Security Project



Quelle: genai.owasp.org

heise academy Videokurse

Neugierig geworden?
Entdecke über 200 Videokurse der heise academy unter **heise-academy.de**



GitHub Actions und Azure Bicep – Infrastructure as Code

In diesem umfassenden Kurs entdeckst du, wie du mit GitHub Actions und Azure Bicep effektiv Infrastructure-as-Code-Strategien (IaC) umsetzen kannst, um deine Infrastrukturmanagementprozesse zu modernisieren und zu automatisieren.

Online verfügbar Mit Tom Wechsler

Mehr erfahren

Azure und Terraform: Einführung in Infrastructure as Code

Dieser Kurs bietet eine umfassende Einführung in die Verwendung von Terraform zur Automatisierung der Azure-Cloud-Infrastruktur. Du lernst, wie du mit dem Azure Resource Manager (ARM) und Terraform die Infrastructure as Code (IaC) definierst, erstellst und verwaltest.

Online verfügbar Mit Tom Wechsler

Mehr erfahren

Sicherheit in Kubernetes

Dieser Kurs vermittelt Methoden zur effektiven Absicherung von Kubernetes-Clustern – von der Vermeidung typischer Fehlkonfigurationen über den Einsatz integrierter Tools bis hin zu externen Werkzeugen, die zusätzlichen Schutz bieten.

Online verfügbar Hubert Ströbitzer

Mehr erfahren

Quellen dieses Whitepapers: EY – Should organisations buy AI systems or build them? (www.ey.com/en_ie/insights/ai/should-organisations-buy-ai-systems-or-build-them) | genai.owasp.org (genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english, genai.owasp.org/llm-top-10, genai.owasp.org/resource/llm-and-gen-ai-data-security-best-practices) | Genesys – Understanding and Managing AI Costs (www.genesys.com/en-gb/blog/post/understanding-and-managing-ai-costs) | LinkedIn: The Cost of Intelligence is Falling Fast (https://www.linkedin.com/posts/robinharbort_the-cost-of-intelligence-is-falling-fast-activity-7320369222075760640-s2kX) | Planview Blog – Build or Buy Generative AI?(blog.planview.com/de/should-you-build-or-buy-your-generative-ai-solution) | resilientcyber.io/p/securitys-ai-driven-dilemma | Rohan Paul – Building vs Buying an LLM (www.rohan-paul.com/p/building-vs-buying-an-llm-key-decision) | saif.google

Wir sind die heise academy

Seit über 40 Jahren steht das renommierte Medienhaus heise für Wissen, Qualität und Unabhängigkeit in der IT. Diese Tradition führen wir fort – mit hochwertiger, digitaler Weiterbildung für alle, die IT lieben.

Unser Ziel: IT-Kompetenzen auf das nächste Level heben.

Der Schlüssel dazu ist kontinuierliche Weiterbildung. So baust du deine IT-Skills aus und meisterst aktuelle Entwicklungen souverän.

Gründe für die heise academy

- Fünf flexible Lernformate – passend für jeden Lerntyp und Zeitplan
- Sieben zentrale IT-Schwerpunkte – von Softwareentwicklung über Security bis KI
- Über 100 IT-Expertinnen und Experten – direkt aus der Praxis

Ob für dich persönlich oder für dein gesamtes Team: Mit der heise academy investierst du in nachhaltiges IT-Know-how – hochwertig, praxisrelevant und immer am Puls der IT.

Zur heise academy

heise-academy.de

Wir sind für dich da!

Dein heise academy Team:

Kontaktiere uns ganz bequem über
unser Kontaktformular, E-Mail oder Telefon.

heise-academy.de/kontakt
lizenzen@heise-academy.de
+49 511 5352 -8602