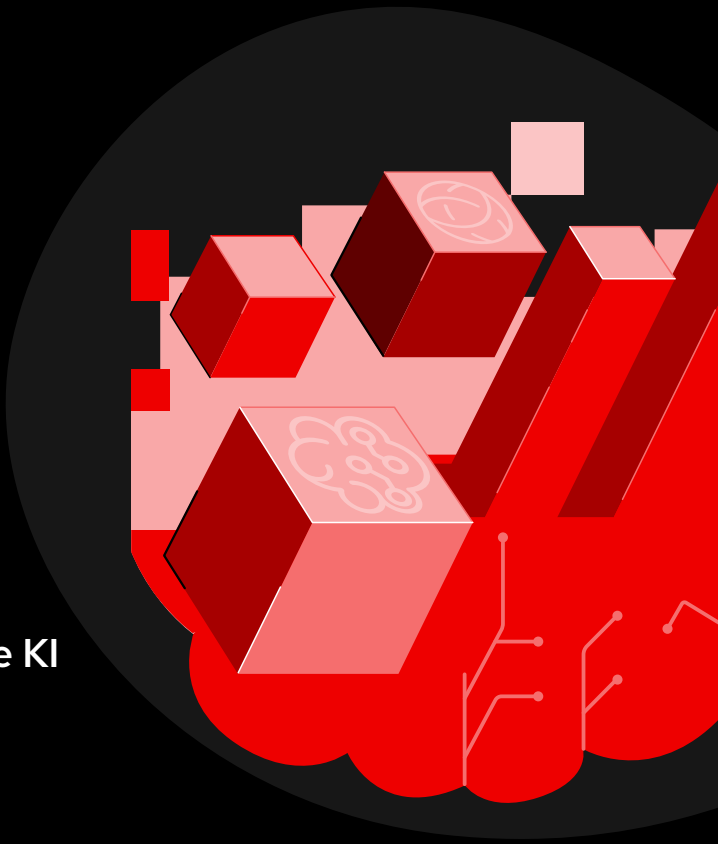
Red Hat

Überlegungen zum

# Aufbau einer Basis für generative KI

# Inhalt



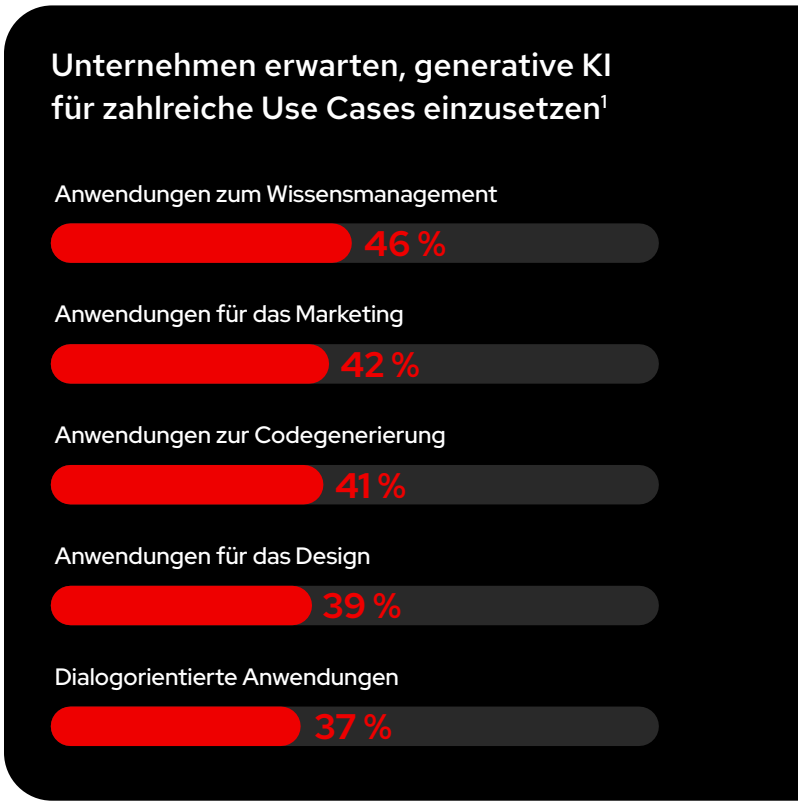


# Neue Chance für geschäftliche Innovation

**Generative künstliche Intelligenz (KI)** ist ein leistungsfähiges Tool für Unternehmen, die innovative Produkte entwickeln, Prozesse optimieren und Wettbewerbsvorteile auf sich schnell verändernden Märkten erzielen wollen. Auf der Basis von Fortschritten im Bereich des Deep Learning und neuronaler Netzwerke geht sie über prädiktive KI-Funktionen hinaus, indem sie nicht nur Daten verarbeitet, sondern auch neue, originelle Inhalte erzeugt. Generative KI verändert die Zusammenarbeit zwischen Mensch und Maschine, inspiriert zu neuen Ansätzen bei der Problemlösung und sorgt branchenübergreifend für erhebliche Geschäftsvorteile.

Weltweit entwickeln Unternehmen neue, innovative Anwendungen mit generativen KI-Technologien. Tatsächlich investieren 39 % derzeit in generative KI-Technologien, während weitere 37 % potenzielle Use Cases prüfen.[1] Hier sind einige der vielen Use Cases für generative KI:

- **Generieren von Prognosen für komplexe Szenarien.** Generative KI kann historische Daten analysieren, Muster erkennen und genaue Prognosen erstellen, die strategische Planung und Risikomanagement unterstützen.
- **Entwickeln von personalisiertem Marketing.** Durch die Datenanalyse zum Erkennen von Kundenpräferenzen und -verhaltensweisen kann generative KI personalisierte Marketingmaterialien – einschließlich E-Mails, Anzeigen und Werbeaktionen – erstellen, die das Engagement und die Konversionsraten maximieren können.
- **Automatisieren und Personalisieren des Kundenservice.** Als Basis für intelligente Chatbots und virtuelle Assistenten kann generative KI automatisch auf Kundenanfragen und -interaktionen reagieren und einen personalisierten, effizienten Kundenservice bieten.




*1 IDC Web Conference Proceeding. „Unlocking Business Success with Generative AI." Dokument Nr. US50789223. Juni 2023.*

# Generative KI wirft neue Fragen auf

Obwohl sich die Vor- und Nachteile der generativen KI erst noch abzeichnen werden, wollen viele Unternehmen jetzt in diese neuen Technologien investieren. Das Wissen um die Probleme in Bezug auf generative KI kann Unternehmen helfen, klare Ethikrichtlinien und Entwicklungsrahmen festzulegen, staatliche und branchenspezifische Vorschriften einzuhalten und potenzielle Probleme zu erkennen und zu beheben.

- **Datenschutz.** Datenschutzbedenken entstehen, wenn generative KI-Modelle mit sensiblen oder personenbezogenen Daten trainiert werden, sodass sich die Frage nach dem Schutz der Privatsphäre des Einzelnen stellt.
- **Dateneigentum.** Die Verwendung von proprietären Modellen – oder von Modellen, die mit proprietären Daten trainiert wurden – führt zu Fragen des Dateneigentums, die zu Rechtsstreitigkeiten führen können.
- **Bias und Fairness.** Die Antworten von generativen KI-Tools spiegeln nachweislich menschliche Vorurteile wider, einschließlich schädlicher Stereotypen und Hassreden.
- **Ethische Nutzung.** Generative KI-Modelle können synthetische Inhalte und Deep Fakes erstellen, die für böswillige Aktivitäten wie Datenschutzverletzungen und Desinformationskampagnen verwendet werden können.
- **Erklärbarkeit und Interpretierbarkeit.** Mangelnde Transparenz bei generativen KI-Tools erschwert die Interpretation, das Verständnis und das Erklären von Modell-Output. Dies führt zu mangelnder Verantwortlichkeit für falsche oder erfundene Informationen.
- **Unbeabsichtigte Konsequenzen.** Die autonome Natur generativer KI kann zu unbeabsichtigten Konsequenzen führen, die Menschen und Organisationen echten Schaden zufügen können.
- **Regulatorische Herausforderungen.** Die rasanten Fortschritte generativer KI-Technologien könnten den Rechtsrahmen überholen, sodass das Erstellen und Durchsetzen von Richtlinien, die für eine verantwortungsvolle und ethische Nutzung sorgen, erschwert wird.
- **Energieverbrauch.** Das Training von KI-Modellen ist rechenintensiv und erfordert viel Energie, wodurch Bedenken hinsichtlich der Umweltauswirkungen und der Nachhaltigkeit aufkommen.

Dieses E-Book behandelt die wichtigsten Überlegungen zum Aufbau einer zuverlässigen Basisinfrastruktur zur Unterstützung generativer KI-Initiativen.

## Vorbereitungen für generative KI

In „Unlocking Business Success with Generative AI" empfiehlt IDC diese Maßnahmen, um Ihr Unternehmen auf generative KI-Initiativen vorzubereiten.[2]

- **Schaffen Sie eine Umgebung für agiles Experimentieren** für priorisierte Use Cases, die Ihren geschäftlichen Anforderungen entsprechen.
- **Entwickeln Sie Unternehmensrichtlinien** für eine verantwortungsvolle Nutzung, die böswilliges Verhalten unterbinden.
- **Bewerten Sie die Auswirkungen** der generativen KI auf die Belegschaft und betreiben Sie proaktives Änderungsmanagement.
- **Arbeiten Sie mit bewährten Technologieanbietern** und Serviceanbietern für Ihre KI-Infrastruktur zusammen.
- **Sichern Sie sich die notwendigen technischen Kompetenzen** durch Neueinstellungen, Trainings oder professionellem Support.

2 *IDC Web Conference Proceeding. „Unlocking Business Success with Generative AI." Dokument Nr. US50789223. Juni 2023.*

# Wichtige Überlegungen zum Aufbau einer Basis für generative KI

Die technologische Basis, die Sie für Ihre generativen KI-Initiativen wählen, kann einen großen Einfluss auf die Akzeptanz und den Gesamterfolg haben. Dieses Kapitel befasst sich mit den wichtigsten Überlegungen für Ihre Basis für generative KI.

## Überlegung 1: Entwickeln mit einem bewährten Toolset

Das Entwickeln von Anwendungen, die auf generativen KI-Modellen basieren, kann sich als komplexe Aufgabe erweisen. Das richtige Toolset – mit Sprachen, Frameworks und Runtimes, die auf Open Source-Projekten und kommerziellen Lösungen basieren – kann die Modellabstimmung beschleunigen und die Anwendungsentwicklung und -bereitstellung vereinfachen.

Wählen Sie eine KI-Basis, die Ihre bevorzugten Toolsets für das schnelle und effiziente Entwickeln innovativer KI-Lösungen bereitstellt. Für eine vereinfachte Zusammenarbeit kann die Unterstützung von explorativer Data Science, Training sowie Tuning durch interaktive Schnittstellen sorgen. Vorintegrierte Toolsets und Self Service-Funktionen unterstützen Sie beim Optimieren von IT-Operationen bei gleichzeitiger Wahrung der Portierbarkeit und Konsistenz in verschiedenen Umgebungen.

## Überlegung 2: Rasches Fine Tuning der Modelle

Da das Trainieren generativer KI-Modelle ein teurer und zeitaufwendiger Prozess ist, entwickeln die meisten Unternehmen KI-Lösungen auf der Grundlage von Basismodellen, die zuvor mit allgemeinen Daten trainiert wurden. Data Scientists verwenden dann verschiedene, domainspezifische Daten, um die Basismodelle für das Ausführen spezieller Aufgaben anzupassen. Fine Tuning kann jedoch immer noch rechenintensiv sein und erfordert leistungsstarke Prozessoren und eine verteilte Hybrid Cloud-Infrastruktur.

Achten Sie auf KI-Plattformen mit verteiltem Workload-Management und Orchestrierungsfunktionen, die Trainingsprozesse – unabhängig von Modellgröße, Datenvolumen oder Dauer – in Hybrid Cloud-Umgebungen bereitstellen. Optionen für das Fine Tuning von Basismodellen in Onsite-Rechenzentren vereinfachen die Compliance mit technischen und regulatorischen Anforderungen für eingeschränkte Modelle. Mit den Batch-Trainingsfunktionen können Sie das Fine Tuning von Workloads vorwegnehmen und die gemeinsame Nutzung und Verwaltung von Ressourcen erleichtern.

### Alternativen zu Fine Tuning-Modellen

Forschungsteams suchen nach Möglichkeiten, Basismodelle schneller und effizienter zu tunen. **Retrieval-augmented generation (RAG)** ist ein KI-Framework zum Abrufen von Fakten aus externen Quellen – wie internen Datenbanken, Firmen-Intranets oder dem Internet – um generative KI-Modelle mit den genauesten und aktuellsten Informationen zu versorgen.

Beim **Prompt Tuning** erhalten KI-Modelle Hinweise oder Front End Prompts (beispielsweise zusätzliche Wörter oder von der KI generierte Zahlen), die das Modell zu einer gewünschten Entscheidung führen, sodass Unternehmen mit begrenzten Daten ein Basismodell auf eine bestimmte Aufgabe abstimmen können.

## Überlegung 3: Effizientes Bedienen von Modellen

Für IT-Operations-Teams kann es eine Herausforderung sein, mit generativen KI-Lösungen außergewöhnliche Benutzererlebnisse zu schaffen. Der variable Bedarf an Anwendungen erfordert eine skalierbare Infrastruktur und eine automatisierte Verwaltung. Für eine effiziente Modellbereitstellung ist die Überwachung der Performance und die schnelle Wiederherstellung früherer Versionen erforderlich. Da KI-Lösungen große Datenmengen verarbeiten, ist die Durchsetzung strenger Sicherheitsstandards in praktisch allen Umgebungen von entscheidender Bedeutung.

Erwägen Sie Plattformen, die generative KI-Modelle und -Anwendungen übergreifend in Hybrid Clouds bereitstellen und skalieren können – einschließlich Onsite-Infrastruktur, Public Cloud-Ressourcen und Edge-Geräten. Optionen zum Bereitstellen generativer KI-Modelle vor Ort oder in isolierten Umgebungen stellen sicher, dass geschützte Daten nicht zum erneuten Trainieren öffentlich verfügbarer Modelle verwendet werden. Auch der Support für Canary-Rollouts und Erklärbarkeits-Tools trägt dazu bei, die Konsistenz und Zuverlässigkeit von Modellantworten zu erhöhen.

## Überlegung 4: Automatisieren des Lifecycle Managements

**CI/CD-Pipelines (Continuous Integration/Continuous Delivery)** können generative KI-Lösungen automatisch bereitstellen und verwalten. Durch erneutes Trainieren und Aktualisieren von Modellen und Anwendungen durch schnelle, schrittweise Änderungen können Sie die Entwicklung beschleunigen und die Modell-Performance erhöhen. KI-Pipelines sind jedoch komplexer als Standard-CI/CD-Workflows, da sie häufig zusätzliche Schritte wie Datenextraktion, Training, Fine Tuning, Validierung und erneutes Training umfassen.

Wählen Sie eine Basis, mit der Sie KI-Pipelines auf Basis von CI/CD-Tools wie Tekton und Jenkins erstellen und in bestehende DevOps-Workflows integrieren können. So können Sie generative KI-Modelle schnell und effizient entwickeln, trainieren, überwachen und erneut trainieren. Mit CD-Tools von **GitOps** können Sie komplexe KI-Lösungen als Code definieren und automatisieren und so für eine konsistente Modell- und Anwendungsbereitstellung sorgen.

### Container für generative KI

**Container-** und **Kubernetes**-Technologien bieten agile Bereitstellung, Verwaltung und Skalierbarkeit, um das cloudnative Entwickeln generativer KI-Lösungen zu beschleunigen. Provisionieren Sie Umgebungen nach Bedarf in Onsite-Rechenzentren, Public Clouds und Edge-Geräten. Erstellen, implementieren, skalieren und verwalten Sie automatisch Container-Instanzen in physischen und virtuellen Infrastrukturen. Integrieren Sie außerdem Komponenten und Datastores aus einem robusten IT-Ökosystem von Open Source- und kommerziellen Anbietern in generative KI-Lösungen. Erfahren Sie mehr über die **Vorteile von Containern für KI**.

# Überlegung 5: Konsistentes Monitoring der Modelle

Generative KI-Modelle können reale, erhebliche Auswirkungen auf Menschen und Unternehmen haben. Durch regelmäßiges Überprüfen des Modellverhaltens können Sie Entscheidungen und Begründungen analysieren, unzureichende Performance erkennen und problematische Verhaltensweisen sofort melden. Eine wirksame Modellsteuerung auf der Basis dieser Informationen trägt dazu bei, dass die Modelle in Produktivumgebungen mit unverfälschten, fairen und korrekten Informationen reagieren.

Achten Sie auf KI-Grundlagen mit zentralisierten Monitoring-Funktionen, die Metriken für Verzerrungen und Datendrift, Anomalieerkennung und punktuelle Erklärbarkeit (per-point explainability) bieten, um Sie beim Untersuchen, Warten und Korrigieren generativer KI-Modelle zu unterstützen. Das kontinuierliche, automatische Monitoring in Produktivumgebungen verbessert die Compliance mit Corporate Model Governance-Standards. Benutzerfreundliche Schnittstellen und leicht verständliche, unkomplizierte Berichte fördern die verantwortungsvolle Nutzung und Wartung der Modelle.

### Wichtige Konzepte bei generativen KI-Modellen

- **Bias** ist das Vorhandensein von Mustern im Modellverhalten, die sich auf die Fairness, Inklusivität und Ethik der erzeugten Ergebnisse auswirken. Dazu gehören das Bevorzugen bestimmter Gruppen oder das Erzeugen von Antworten, die mit Stereotypen übereinstimmen.
- **Datendrift** tritt auf, wenn sich die statistischen Eigenschaften der Trainingsdaten im Laufe der Zeit ändern. Dies führt zu einer Abnahme der Modell-Performance und zum Generieren von ungenaueren oder weniger relevanten Antworten.
- **Bei** der Erkennung von Anomalien geht es um das Identifizieren und Melden von Modellverhalten, das ungewöhnlich ist oder von den beim Training beobachteten Beispielen abweicht.
- **Die punktuelle Erklärbarkeit** ist die Fähigkeit, zu verstehen, warum Modelle bestimmte Ergebnisse erzeugen, und bietet Klarheit für Anwendungen, bei denen Transparenz entscheidend ist.

# Überlegung 6: Nutzen der Vorteile von Partnernetzwerken

Generative KI-Lösungen erfordern mehrere integrierte Komponenten, um innovative Benutzererlebnisse zu bieten. Mit der richtigen Kombination von Technologien aus einem gemeinsamen IT-Ökosystem zuverlässiger Anbieter können Sie die Anwendungsentwicklung beschleunigen, Probleme mit Verzerrungen und Datendrift lösen und eine konsistente, zuverlässige Performance für Ihre gesamte Lösung sicherstellen.

Suchen Sie nach Plattformanbietern mit umfangreichen, zertifizierten Partnernetzwerken, die Komplettlösungen für das Entwickeln und Bereitstellen generativer KI-Modelle und -Anwendungen anbieten. Eine große Auswahl an Komponenten, von der Datenintegration und -aufbereitung bis hin zum Modelltraining und -service, unterstützt Sie beim schnelleren und effizienteren Entwickeln und Bereitstellen von KI-Lösungen. Wenn Sie sich für zertifizierte Lösungen mit bewährter Interoperabilität entscheiden, können Sie IT-Supportanfragen reduzieren und die Produktivität steigern.

# Überlegung 7: Zusammenarbeit mit Fachleuten für Plattformen

Die effektive Bereitstellung und Verwaltung generativer KI-Lösungen erfordert spezialisiertes Wissen und Erfahrung. Die Anforderungen an die Skalierbarkeit, die Zuverlässigkeit und die Integration in bestehende Systeme können das Deployment in der Produktion erschweren. Eine effiziente Nutzung von Rechenressourcen kann zu unnötigen Kosten führen. Zudem kann die Nichteinhaltung von Sicherheitsstandards, Datenschutzrichtlinien und gesetzlichen Rahmenbedingungen für KI zu unbeabsichtigten Folgen führen.

Wählen Sie Anbieter mit Expertenteams, die umfassende Unterstützung und Anleitung für das Entwickeln generativer KI-Lösungen bieten. So können beispielsweise engagierte Engineers die gesamte Plattform mit den Tools, Ressourcen und Kenntnissen unterstützen, um Ihre KI-Projekte zu beschleunigen. Fachkundige Consultants können Herausforderungen beim Deployment lösen, die Effizienz der Infrastruktur optimieren und für die Interoperabilität Ihrer KI-Lösung sorgen. Dazu können professionelle Trainingsservices Sie beim Erwerb von Wissen und Know-how unterstützen, damit Sie schneller mit neuen generativen KI-Projekten beginnen können.

## Generative KI erfordert Zusammenarbeit

Der Aufbau eines Teams mit einer Vielzahl von Fähigkeiten ist für den Erfolg generativer KI-Projekte entscheidend.[3]

- **Business Leader** vertreten die Menschen, die diese Lösung nutzen oder von ihr betroffen sind.
- **KI-Fachleute** tunen, warten und aktualisieren generative KI-Modelle.
- **Data Scientists bereiten die Daten vor** und stellen korrekte, unverzerrte Trainingsdaten für Modelle bereit.
- **Beauftragte für Ethik und Compliance stellen sicher,** dass generative KI-Initiativen den Vorschriften entsprechen.
- **IT-Operations-Specialists** integrieren Lösungen in die bestehende Infrastruktur und setzen Sicherheitsrichtlinien durch.

*3 Kearney. „**Standing up tiger teams to tackle generative AI complexity**," November 2023.*

# Schnelles Innovieren mit einer flexiblen und offenen Basis

Red Hat bietet ein komplettes Technologieportfolio, bewährtes Fachwissen und strategische Partnerschaften, damit Sie Ihre Ziele für generative KI erreichen können. Wir bieten eine Basis für das Entwickeln und Bereitstellen generativer KI-Modelle und -Anwendungen sowie Services und Training für eine schnelle Einführung.

**Red Hat® OpenShift®** ist eine einheitliche, unternehmensgerechte Anwendungsplattform für cloudnative Innovationen. On-Demand-Rechenressourcen, Unterstützung für Hardware-Beschleunigung sowie Konsistenz in Onsite-, Public Cloud- und Edge-Umgebungen sorgen für die Geschwindigkeit und Flexibilität, die Teams für ihren Erfolg benötigen. Mit Red Hat OpenShift können Sie eine Self Service-Plattform für Data Scientists, Data Engineers und Entwicklungsteams schaffen, um schnell intelligente Anwendungen zu entwickeln. Dank der Funktionen für die Zusammenarbeit können Teams containerisierte Modellierungsergebnisse erstellen und mit Peers und Mitgliedern des Entwicklungsteams auf konsistente Weise austauschen.

**Red Hat OpenShift AI** basiert auf Red Hat OpenShift und bietet eine umfassende Plattform für das Entwickeln, Trainieren, Fine Tuning, Bereitstellen und Monitoring von Modellen und Anwendungen. Gleichzeitig erfüllt die Lösung die Workload- und Performance-Anforderungen moderner generativer KI-Lösungen. In einer kollaborativen, konsistenten Umgebung, die Angebote wichtiger, zertifizierter Partner wie NVIDIA, Intel, Starburst, Anaconda, IBM, Run:ai und Pachyderm integriert, können Teams schnell vom Experiment zur Produktion übergehen. Zusammen mit unserem Technologienetzwerk bietet Red Hat OpenShift AI Komponenten und Funktionen, die das Entwickeln und Bereitstellen innovativer, generativer KI-Lösungen in Hybrid Clouds beschleunigen.

**IBM watsonx.ai AI Studio** bietet eine Auswahl an Modellen und Deployment-Optionen mit den generativen KI-Funktionen, die Sie für Ihre intelligenten Anwendungen nutzen können. Stellen Sie Modelle – einschließlich Open Source-, Drittanbieter- und von IBM entwickelte Basismodelle – an den Standorten Ihrer Workloads bereit, um die Performance und Effizienz Ihrer KI-Lösungen zu steigern. Dank der von **IBM entwickelten Basismodelle**, die auf unternehmensrelevanten Daten trainiert wurden, verstehen Ihre generativen KI-Lösungen die Feinheiten Ihres Unternehmensbereichs und verschaffen Ihnen einen Wettbewerbsvorteil.

**Red Hat® Ansible® Lightspeed mit IBM watsonx Code Assistant** ist ein generativer KI-Service, der Ihren Teams hilft, Automatisierungsinhalte effizienter zu erstellen, einzuführen und zu verwalten. Red Hat Ansible Lightspeed ist mit IBM watsonx Code Assistant verbunden und hilft Ihnen, Ihre Vorstellungen zur Automatisierung in Ansible-Code mit Eingaben in natürlicher Sprache umzusetzen. So können Sie die Produktivität steigern und den Zugang zur Automatisierung in Ihrem Unternehmen erleichtern.

# Bereit für generative KI?

## Generative KI ist ein leistungsstarkes Tool für das Erstellen origineller Inhalte und verändert die Art und Weise, wie wir mit Anwendungen und Technologien interagieren.

Durch Technologie, Expertise und Partnerschaften bietet RedHat die Basis für Ihre Teams, um KI-Anwendungen und ML-Modelle mit Transparenz und Kontrolle zu entwickeln und bereitzustellen. Wir verwenden sogar unsere eigenen KI-Tools und Plattformen, um die Nützlichkeit anderer Open Source-Software zu verbessern. Außerdem bieten wir Ihnen durch unsere Partnerintegrationen Zugang zu einem IT-Ökosystem zuverlässiger KI-Tools, die für Open Source-Plattformen wie Red Hat OpenShift AI entwickelt wurden.

**Erfahren Sie mehr und testen Sie Red Hat OpenShift AI kostenlos.**

### Schneller Einstieg mit Red Hat Consulting

Arbeiten Sie mit den Fachleuten von Red Hat zusammen, um Ihre KI/ML-Projekte voranzutreiben. Red Hat bietet Services für Consulting und Training, damit Ihr Unternehmen KI/ML schneller einführen kann.

- Mehr über KI/ML-Services erfahren: red.ht/aiml-consulting
- Vereinbaren Sie eine kostenlose Discovery Session: redhat.com/consulting



1057715_0324_KVM

Red Hat