Ein Leitfaden für Führungskräfte zur Operationalisierung von generativer KI

Von der Erprobung zur Umsetzung: So können Sie generative KI in realen Szenarien einsetzen

elastic | The Search AI Company

elastic.co/de | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Inhaltsverzeichnis

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Beginnen Sie Ihre Reise zur generativen KI

Generative KI war 2023 die disruptivste Technologie. In allen Branchen lautet die Vorhersage, dass generative KI in den nächsten Jahren zum umfassend prägenden Faktor wird. Dennoch – wer kann von sich behaupten, den Code geknackt zu haben und zu wissen, wie sich die Technologie jetzt einsetzen lässt?

Während die meisten Unternehmen gerade erst beginnen, sich mit generativer KI auseinanderzusetzen, können einige bereits erste Ergebnisse vorweisen. Beispielsweise können die Support-Experten von Cisco sofort eine Zusammenfassung relevanter Antworten von ähnlichen Supportfällen, internen Diskussionsforen oder Knowledge-Artikeln zu bestimmten Kundenproblemen finden. Cisco konnte bereits von den Vorteilen der generativen KI profitieren und 90 % der Supportanfragen mit seinem neu konzipierten Suchangebot lösen. Dadurch konnten pro Monat 5.000 Stunden für Supportmitarbeiter eingespart werden.[1]

Im Bereich E-Commerce hatten Sie es wahrscheinlich auch schon mit generativer KI zu tun. Generative KI kann die früheren Einkäufe eines Kunden analysieren und den Verlauf sowie die Einstellungen durchsuchen, um personalisierte Produktempfehlungen mit einem Chatbot zu generieren. Im Backend-Bereich verspricht die Nutzung von generativer KI eine Stärkung von Kundenengagement und -bindung, verbesserte Betrugserkennung und vieles mehr.

Um die Fähigkeiten der generativen KI zu entmystifizieren und über ihren Einsatz in Ihrer Umgebung zu entscheiden, brauchen Sie eine Schritt-für-Schritt-Anleitung zur Aktivierung Ihrer Daten. Mit diesem E-Book begleiten wir Sie bei Ihrer Reise vom **Wunschdenken zum KI-Experten.** Stellen Sie es sich wie eine Roadmap vor, die Sie dabei unterstützt, Ihre Geschäftsergebnisse mithilfe von generativer KI zu revolutionieren.

99 % *der Unternehmen davon überzeugt, dass generative KI das Potenzial hat, den Wandel in ihrem Unternehmen intern oder extern voranzutreiben*[2]

Dennoch sind nur

32 % *der Führungskräfte haben Vertrauen in ihre Fähigkeit, KI in ihren Unternehmen zu implementieren*[3]

[1] Elastic, Cisco erstellt KI-gestützte Sucherlebnisse mit Elastic auf Google Cloud 2024
[2] Elastic, The Elastic Generative AI Report, (2024)
[3] Russell Reynolds, Embracing the Unknown: How Leaders Engage with Generative AI in the Face of Uncertainty, (2024).

Von diesem Leitfaden können Sie Folgendes erwarten:

Schritt 0

Stellen Sie sich die Frage, wie Ihnen generative KI helfen kann. Was möchten Sie mit generativer KI erreichen?

Schritt 1

Identifizieren Sie Ihre idealen Ergebnisse. Wie sieht eine erfolgreiche Implementierung für Ihre Anwendungsfälle aus?

Schritt 2

Ermitteln Sie die Auswirkungen und messen Sie den Erfolg. Überlegen Sie, welche organisatorischen Prozesse betroffen sind und wie dies geschieht.

Schritt 3

Wählen Sie eine Implementierungs-strategie. Erkunden Sie Ihre Optionen.

Schritt 4

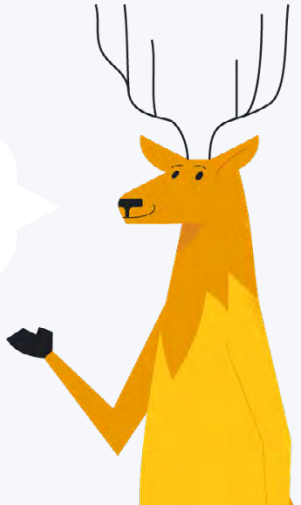
Starten Sie einen Test und verfolgen Sie ein iteratives Herangehen.

Schritt 5

Legen Sie Governance-Standards fest und berücksichtigen Sie Datensicher-heitsbedenken.

Schritt 6

Legen Sie eine Zeitleiste mit Zielvorgaben für Ihr Team fest.



Lassen Sie uns zunächst die Grundlagen auffrischen.

elastic.co/de | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Teil 1: Generative KI – ein allgemeiner Überblick

Sie müssen kein Experte für generative KI sein, um einen Plan für ihre Operationalisierung zu erstellen. Wenn Sie jedoch verstehen, welche Komponenten im Spiel sind, können Sie während des gesamten Prozesses informierte und strategische Entscheidungen treffen. Lassen Sie uns die Bausteine beleuchten.

Was ist generative KI?

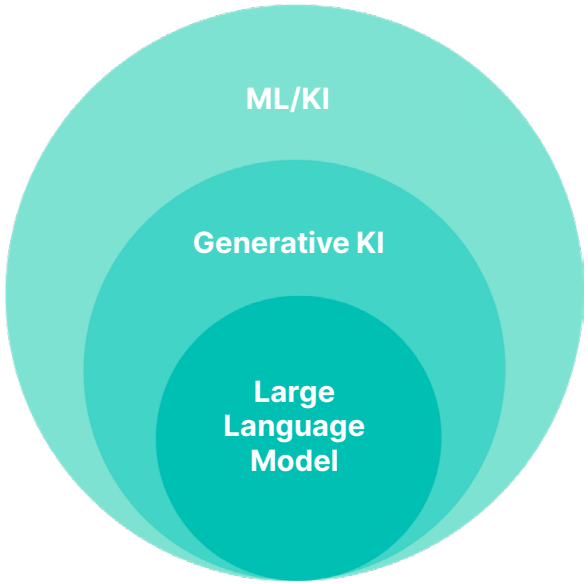
Generative künstliche Intelligenz, oder generative KI, ist eine Bezeichnung für Deep-Learning-Modelle, die auf Aufforderung Outputs generieren können. Es ist wichtig, zu verstehen, dass die generative Fähigkeit dieser Technologie darauf beruht, dass sie mithilfe von Machine Learning statistisch wahrscheinliche Ergebnisse vorhersagen kann. **Daten sind das Herzstück bei der Operationalisierung von generativer KI und der Schlüssel zum Erfolg – für die Implementierung ebenso wie für die Ergebnisse.** Mehr darüber in Kürze.

Was ist Machine Learning?

Machine Learning oder ML ist eine Untergruppe der künstlichen Intelligenz, die Algorithmen verwendet, um aus Daten Wissen zu gewinnen. Diese Algorithmen analysieren Daten und „lernen" – in einem überwachten, halbüberwachten oder unüberwachten Kontext – die Muster und Ähnlichkeiten, mit deren Hilfe sie Entscheidungen treffen können. Machine Learning ist die zugrundeliegende Technologie, die generativer KI, wie z. b. großen Sprachmodellen, die Fähigkeit zum kontinuierlichen „Lernen" verleiht.

Was ist ein großes Sprachmodell (Large Language Model, LLM)?

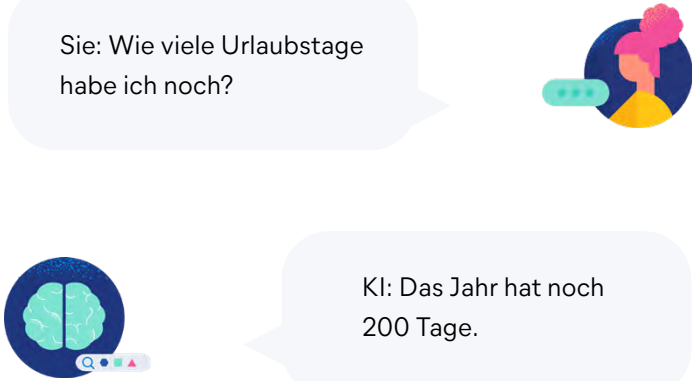
Ein großes Sprachmodell oder LLM ist ein Computermodell, das hinter dem Machine Learning steckt. Es ist eine Art von generativer KI, die sich speziell mit menschlicher Sprache beschäftigt. Da ein LLM mit umfangreichen, hauptsächlich öffentlichen Sprachdatensätzen trainiert wurde, kann es eine Vielzahl von Aufgaben zur Verarbeitung natürlicher Sprache (NLP) ausführen, darunter das Erkennen, Analysieren, Zusammenfassen, Vorhersagen, Übersetzen oder Generieren von Text. Im Kontext der Operationalisierung von generativer KI sind es die LLMs, denen die generative KI die Fähigkeit verdankt, in natürlicher (oder menschlicher) Sprache zu kommunizieren.



 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Wenden wir uns dem Thema Halluzinationen zu

Eine Halluzination ist ein falsches oder irreführendes Ergebnis, das von einem LLM generiert wird. Sie haben wahrscheinlich schon von den manchmal fragwürdigen Antworten von ChatGPT gehört. Der Output erscheint seriös ... aber ist er es wirklich? Wenn das LLM (ChatGPT basiert auf einem LLM) keine Antwort findet, erfindet es eine. Diesen wunden Punkt muss man berücksichtigen, wenn man den Einsatz von LLMS in Unternehmensanwendungen erwägt. Wie stellen Sie sicher, dass die generierten Outputs relevant und richtig sind? An dieser Stelle komm Retrieval Augmented Generation (RAG) ins Spiel.

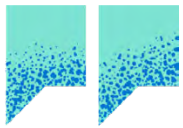


 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Was versteht man unter Retrieval Augmented Generation (RAG)?

Stellen Sie sich Retrieval Augmented Generation oder RAG wie eine Verteidigungslinie gegen Halluzinationen vor. Der von einem LLM erzeugte Output wird durch den Abruf von Informationen aus einem bestimmten Datensatz oder Datenkontext, den Sie mit Hilfe einer hochrelevanten Suche auf der Grundlage einer Vektordatenbank bereitstellen, ergänzt oder „geprüft“. Beispielsweise durchsucht ein Unternehmen mittels RAG und als Antwort auf eine Benutzeranfrage seine Richtliniendokumente und stellt einem LLM relevante Antworten zur Verfügung, so dass es unter Verwendung der Unternehmensrichtlinien antworten kann. Aber RAG dient nicht nur als Verteidigung gegen Halluzinationen, **sondern bietet Ihnen auch die Möglichkeit, proprietäre Datensätze mit der generativen KI zu verwenden.** Das ist der größte Vorteil von RAG.

Im Kontext der Operationalisierung von generativer KI für Unternehmensanwendungen ist RAG aus mehreren Gründen wichtig: Es kann bessere, relevantere Ergebnisse liefern und bietet eine schnelle Möglichkeit, eigene Daten zu booten oder zu nutzen. Es ist auch kosteneffizienter als das Trainieren oder Erstellen eines eigenen LLM. Somit ist RAG der Schlüssel zur erfolgreichen Integration von generativer KI. RAG geht über die Grenzen allgemeiner LLMs hinaus, um „die Suchmaschine der nächsten Generation“ zu erstellen.



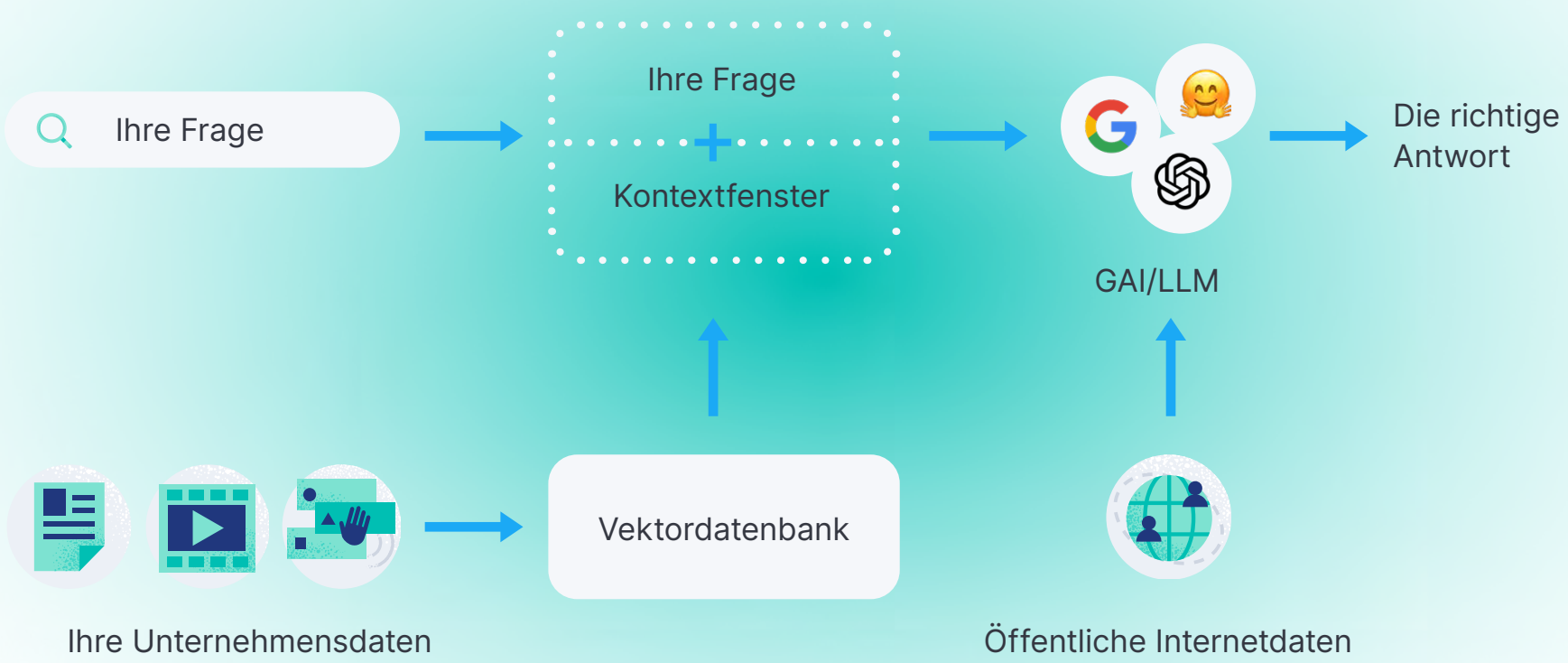
„Mit RAG können Sie die Suchmaschine der nächsten Generation erstellen.“

Baha Azarmi
VP, Global Customer Engineering bei Elastic

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Retrieval Augmented Generation (RAG)

Mit RAG können Sie generative KI mit Ihren proprietären Datensätzen verwenden.



 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

RAG bietet eine neue Möglichkeit zur Beantwortung von Nutzerfragen:

Normale Suche

Eine Person sucht nach **einem Begriff**

Richtlinie zum Homeoffice

Eine Abfrage wird durchgeführt

```
query = {
“bool”: {
“should”: [
{
“text_expansion”:{
“ml.inference.text_expanded_predicted_
value”:{
“model_id”:model_id.
“model_text”:question
}
```

Ergebnisse werden präsentiert

	Dokumententitel	Datum hinzugefügt
1	Verhaltenskodex für Mitarbeiter	01/01/2010
2	Richtlinie zur IT-Nutzung	01/01/2015
3	Startseite in Arbeit	01/01/2022
	etc...etc...	

Nutzer wählen ein Dokument aus und lesen seinen Inhalt

Verhaltenskodex für Mitarbeiter

Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum Lorem ipsum
Lorem ipsum

Generative KI ohne RAG

Eine Person stellt **eine Frage**

Was ist unsere Richtlinie zum Homeoffice?

Eine Abfrage wird durchgeführt

```
query = {
“bool”: {
“should”: [
{
“text_expansion”:{
“ml.inference.text_expanded_predicted_
value”:{
“model_id”:model_id.
“model_text”:question
}
```

Es wird eine Antwort angezeigt, die keinen Zusammenhang mit Ihrem Bereich hat

Eine Richtlinie zum Homeoffice ist erforderlich, wenn Sie Mitarbeiter haben, die hybrid arbeiten oder ....

RAG-fähig

Eine Person stellt **eine Frage**

Was ist unsere Richtlinie zum Homeoffice?

Eine Abfrage wird durchgeführt

```
query = {
“bool”: {
“should”: [
{
“text_expansion”:{
“ml.inference.text_expanded_predicted_
value”:{
“model_id”:model_id.
“model_text”:question
}
}
]
{
“match”:{
“text”: question
}
```

Ergebnisse werden als Kontext angezeigt

	Dokumententitel	Datum hinzugefügt
1	Verhaltenskodex für Mitarbeiter	01/01/2010
2	Richtlinie zur IT-Nutzung	01/01/2015
3	Startseite in Arbeit	01/01/2022
	etc...etc...	

Aus den Suchergebnissen eines LLM wird eine Antwort hergeleitet

Mitarbeiter werden dazu angehalten, zuhause zu arbeiten, wenn sie dort effektiv ... etc...etc...

Was ist eine Vektordatenbank?

Eine Vektordatenbank speichert Vektoreinbettungen oder numerische Darstellungen von Wörtern, Bildern oder Video. Diese Einbettungen sind mehrdimensional und ermöglichen eine semantische Suche. Bei dieser Art der Suche wird nach der Absicht und der kontextuellen Bedeutung einer Abfrage gesucht. Im Gegensatz dazu wird bei einer textuellen Suche nur nach Ergebnissen gesucht, die mit den Suchbegriffen in der Abfrage übereinstimmen.

Im Kontext von RAG ermöglicht eine Vektordatenbank eine schnelle semantische Suche auf der Grundlage der Eingabeaufforderung an die generative KI. Das ist es, was RAG ermöglicht.

Generative KI ist zwar gut bei NLP, mit der herkömmlichen Schlüsselwortsuche ist es jedoch nicht möglich, natürliche Sprache zu verstehen und die besten Ergebnisse zur Weitergabe an die generative KI zu liefern. Dagegen kann die generative KI mit einer Vektordatenbank, die sie mit Suchergebnissen beliefert, die semantisch der ursprünglichen Eingabeaufforderung ähneln, relevantere Antworten generieren. Stellen Sie sich eine Vektordatenbank als die Wissensbank vor, die es der generativen KI ermöglicht, Fragen mit genauen Informationen zu beantworten.

Generative KI beschränkt sich jedoch nicht auf Vektordatenbanken. Mithilfe von RAG kann generative KI auf relationale Datenbanken, Graphdatenbanken, dokumentenbasierte Datenbanken oder sogar Suchmaschinen für Schlüsselwörter zugreifen. Welche Datenbank für Sie die beste ist, hängt häufig von der Art der Daten, den jeweils verwendeten Algorithmen und den Leistungsanforderungen des Systems ab. Beispielsweise können relationale Datenbanken für strukturierte Daten verwendet werden, während sich Graphdatenbanken gut für Daten mit komplexen Beziehungen und herkömmliche Suchmaschinen für die Volltextsuche eignen.

„Alle Wege führen zur hybriden Suche."

Serena Chou, Director, Product Management bei Elastic

Während die **semantische Suche** Ergebnisse erbringt, die mit der Bedeutung einer Abfrage übereinstimmen, spielt die **Schlüsselwortsuche** immer noch eine wichtige Rolle beim Abgleich der Ergebnisse mit exakten Suchbegriffen aus den Abfragen. Hybride Suche bezeichnet ein Vorgehen, bei dem sowohl die Vektorsuche – häufig für die semantische Suche verwendet – als auch Technologien für die Schlüsselwortsuche eingesetzt werden, um einer generativen KI die relevantesten Ergebnisse zu liefern.

Fazit: Eine hybride Suchlösung erbringt wahrscheinlich die relevantesten Ergebnisse für Erlebnisse der generativen KI in Ihrem Unternehmen.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Welche Möglichkeiten eröffnet generative KI?

Wir haben nun viel über die zugrundeliegenden Technologien und Grundkonzepte gesprochen, aber was genau kann generative KI leisten?

Erstellen

Indem generative KI mit ihren Trainingsdaten Muster lernt, kann sie Outputs „erstellen“ oder *generieren*. Es wird ein iterativer Prozess an vorhandenen Daten ausgeführt, um neue Ideen, Bilder, Erkenntnisse und mehr zu gewinnen.

Zusammenfassen

Dank ihrer Fähigkeiten zur Verarbeitung natürlicher Sprache kann generative KI einen Text analysieren und zusammenfassen. Müssen Sie lange Dokumente in einer kurzen Zeit durchlesen? Hier schafft generative KI Abhilfe.

Entdecken

Der Schlüssel zu generativer KI ist die Suchtechnologie, auf der sie basiert. Sie ermöglicht dem Tool, das generative KI einsetzt, eine Abfrage zu erhalten, eine riesige Menge an privaten oder öffentlichen Daten zu durchsuchen und eine Antwort zu generieren.

Automatisieren

Nehmen wir an, dass Ihr Unternehmen zwei unterschiedliche Cloud-Plattformen für verschiedene Dienste verwendet. Jede Cloud-Plattform generiert Logdaten in unterschiedlichen Formaten. Durch die Umwandlung dieser Daten in dasselbe Format und ihre Zuordnung zur KI kann Ihr Team sie zusammenfassen und mithilfe der generativen KI Fragen zu den Daten stellen. Ihr IT-Team kann sich auf die Überwachung und Verwaltung Ihrer Systeme fokussieren, anstatt arbeitsintensive Aufgaben auszuführen.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

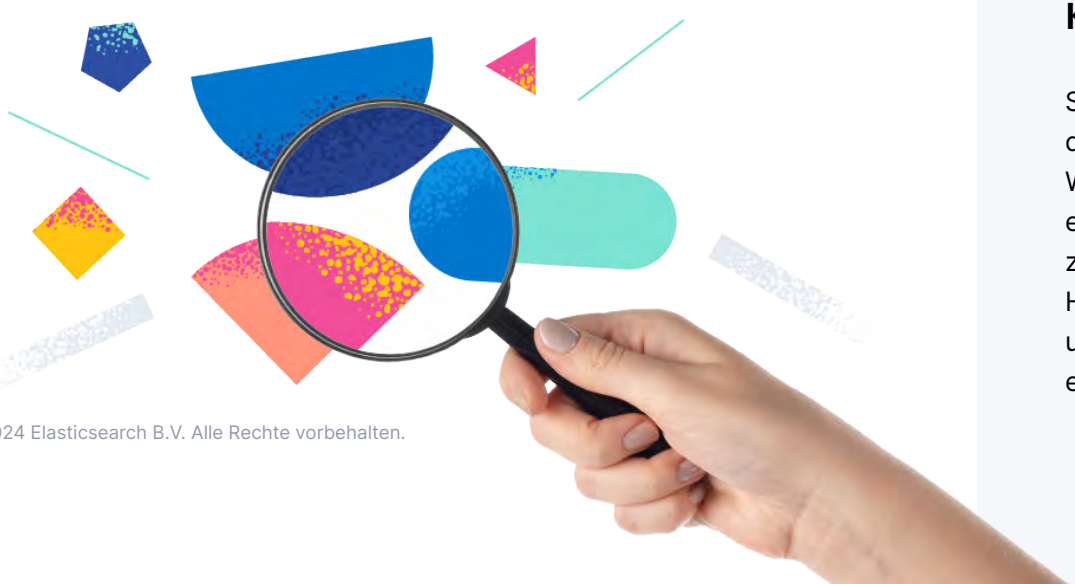
Schritt 0: Ermitteln von Hauptanreiz und Potenzial

Mit so vielen Möglichkeiten der Wertschöpfung ist es entscheidend, einen Ausgangspunkt festzulegen. Wie können Sie die Kraft generativer KI nutzen, um Ihr Team zu stärken, die sich schnell verändernden Erwartungen der Kunden zu erfüllen und Ihr Unternehmen zu neuen Höhen zu führen? Indem Sie sich auf den einen Bereich konzentrieren, in dem generative KI Ihrem Unternehmen den größten Wert bieten kann.

Stellen Sie sich folgende Fragen:

1 Welches Problem möchte ich lösen?

Gibt es einen Bereich, in dem Ihr Unternehmen besonders ineffizient ist? Wo wenden Ihre Beschäftigten viel Zeit für sich wiederholende Aufgaben auf? Sind sie ständig damit beschäftigt, nach vorhandenen Informationen in internen Datenbanken oder externen Engines zu suchen?



Dazu die folgenden Beispiele:

Haben Ihre Beschäftigten Schwierigkeiten, Informationen zu finden – egal ob es sich um Projekt-Updates oder HR-bezogene Informationen handelt? Hat Ihr Sicherheitsteam Aufgaben, die es automatisieren kann, um mehr Zeit für proaktives Handeln zu gewinnen? Haben Sie bei den Prozessen Ihres Kundenservices mit Entropie zu kämpfen, weil Ihre Kundenservice-Techniker bekannte Probleme und aktuelle Korrekturen nicht immer in Echtzeit an die Kundenservice-Mitarbeiter weitergeben?

Kleine Erinnerung:

Sie müssen mit dem Team arbeiten, dessen Workflow Sie beeinflussen. Wenn Sie beispielsweise entscheiden, einen ineffizienten HR-Prozess zu aktualisieren, wird es wichtig sein, HR von Beginn an einzubeziehen, um die Zustimmung der Stakeholder einzuholen und den Support zu stärken.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

2 Handelt es sich um ein Problem, das ich mit einer Wissensdatenbank lösen kann?

Eine Wissensdatenbank ist eine Sammlung informativer Inhalte – ob Support-Artikel oder interne Prozesse.

Überlegen Sie, auf welchen Content Sie sich stützen und ob Sie in diesem Content die Antwort darauf finden, wie Sie Ihr Problem effizienter lösen können. Reicht er für eine Automatisierung und Personalisierung der Antworten aus?

Dazu ein Beispiel: Sie haben herausgefunden, dass Beschäftigte relativ viel Arbeitszeit damit verbringen, nach HR-bezogenen Informationen zu suchen. In Ihrem Team gibt es keinen HR-Experten, darum werden die Beschäftigten an das Intranet des Unternehmens verwiesen, wo sie eine Vielzahl von Richtliniendokumenten durcharbeiten müssen, um beispielsweise herauszufinden, wie viele Urlaubstage ihnen für dieses Jahr noch bleiben. Um dieses Problem zu lösen, benötigen Sie eine Wissensdatenbank, wie beispielsweise Ihre HR-Richtliniendokumente, sowie Zugriff auf die personenbezogenen Daten der Beschäftigten, um die Antwort zu personalisieren.

„Wenn Sie an einem Faden ziehen, kommen Tausende weitere [Fäden] zum Vorschein."

Baha Azarmi, VP, Global Customer Engineering bei Elastic

Schritt 0 besteht darin, zu verstehen, welche Prozesse die Produktivität einschränken. Indem der Zeitaufwand für alltägliche Aufgaben reduziert wird, haben Ihre Beschäftigten mehr Kapazitäten, um kreativ zu sein. Eine gut ausgeführte Implementierung ist eine verantwortungsvolle. **Investitionen in Ihre Beschäftigten, deren Weiterbildung und die Planung von angepassten Arbeitsabläufen und Prozessen sind entscheidende Bestandteile für einen erfolgreichen Einsatz von generativer KI.**

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Berücksichtigen Sie bei der Suche nach den Gründen das folgende Prinzip: **K.I.S.S. – keep it simple and specific.** Zu identifizieren, welchen Anwendungsfall für generative KI Sie zuerst in Angriff nehmen möchten, ist ein wichtiger erster Schritt bei der Operationalisierung generativer KI. Außerdem können Sie mit kleineren Projekten eine effektive Implementierung erreichen.

Im vorherigen HR-Szenario könnte generative KI beispielsweise für mehrere Anwendungsfälle zum Einsatz kommen:

1

Discovery

Eine Mitarbeiterin gibt in die Oberfläche folgende Frage ein: Wie viele Urlaubstage habe ich dieses Jahr noch? Zur Beantwortung der Abfrage muss die KI eine Suche durchführen und die für die Abfrage relevanten Dokumente anzeigen, z. B. die Dokumente der HR-Richtlinie und die Personalakte der Mitarbeiterin.

2

Zusammenfassung

Im nächsten Schritt analysiert die generative KI eventuell die Dokumente und fasst sie für die Mitarbeiterin in einer Gesprächsantwort zusammen. „Sie haben in diesem Jahr noch 10 Urlaubstage und vier bewegliche Ferientage, die Sie nutzen können. Auf der Intranetseite finden Sie mehr Richtlinien zum Urlaubsanspruch.“

3

Erstellung und Automatisierung

Der Chatbot könnte Führungskräften Zeit ersparen, indem er eine Antwort zur Genehmigung oder Ablehnung eines Urlaubsantrags erstellt und sogar eine Begründung dafür liefert. Er könnte auch Kalender-Einladungen erstellen und die Urlaubsanfrage in das System eingeben.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Anpassung an Ihre Branche

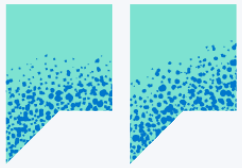
Copilot, Assistent, Bot – generative KI nimmt verschiedene „Formen“ an, die wertvolle Dienste zur Produktivitätssteigerung in einer Vielzahl von Bereichen und Branchen bieten. Als Copilot für Sicherheit und Observability oder mit internen und externen Anwendungen kann generative KI Unternehmen helfen, ihre Effizienz zu steigern, ihre Sicherheitsmaßnahmen zu verstärken, das Kundenerlebnis zu verbessern und sich schneller von Wettbewerbern abzuheben.

Indem sie Daten nutzen, um die besten KI-Antworten zu finden und die Leistungsfähigkeit der zugrunde liegenden Suchtechnologie zu nutzen, **können Unternehmen den Zeitaufwand für alltägliche Aufgaben verringern, die Reaktionszeiten verkürzen und die Gesamtproduktivität steigern**. Wenn dann noch RAG ins Spiel kommt, können Sie Ihre proprietären Daten in die generative KI einfließen lassen, um Antworten zu generieren, die sicher sind und gleichzeitig die Berechtigungen auf Dokumenten- und Benutzerebene respektieren.

Plötzlich haben Sie die nächste Stufe an Geschwindigkeit und Relevanz erreicht – die gleiche Geschwindigkeit und Relevanz, die zunehmend technikaffine Kunden erwarten. Von entscheidender Bedeutung ist, dass Ihre Dienste den Erwartungen entsprechen. Dasselbe gilt für die Nutzung von generativer KI in den Bereichen, die für das Benutzererlebnis am wichtigsten sind. Es gibt nichts Schlimmeres, als in hochmoderne Ressourcen zu investieren, die niemandem nutzen.

Generative KI wird meistens in die IT-Infrastruktur eines Unternehmens als KI-Assistent oder Copilot für Security und/oder Observability integriert.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.



Jedem Unternehmen eröffnen sich Chancen durch generative KI, weil generative KI grundsätzlich eine viel menschlichere und intuitivere Art ist, Informationen aus Informationssystemen zu gewinnen.

Ash Kulkarni
CEO bei Elastic

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

KI-Assistenten

Holen Sie mit internen und externen Apps das Beste aus den Konversationsfähigkeiten der generativen KI für Beschäftigte und Kunden heraus. KI-Assistenten bieten allen Nutzern flexible, anpassungsfähige und persönliche Hilfe, indem sie als praktische Experten, persönliche Einkäufer oder sogar Terminerinnerungen fungieren.

Copilot für Security und Observability

Steigern Sie Ihre Observability- und Security-Fähigkeiten mit Copiloten für generative KI. Die Copiloten für generative KI sind auf die Zusammenarbeit mit IT-Teams ausgelegt und fungieren als kompetente Partner bei der Problemlösung. Beispielsweise können Sie Ihren Copiloten nach einer detaillierten Beschreibung fragen, warum eine Sicherheitswarnung ausgelöst wurde. Er empfiehlt Ihnen daraufhin Schritte zur Eingrenzung und Behebung des Angriffs (auf der Grundlage früherer ähnlicher Angriffe, denen Ihr Unternehmen ausgesetzt war). Diese Art der Eingabeaufforderung kann ein dynamisches Runbook für die Organisation generieren.

Dank dieser Integrationen können Unternehmen sämtlicher Branchen ihr Potenzial für Personalisierung, Automatisierung und Produktivität steigern, wodurch sich **drei große Anwendungsfälle für generative KI ergeben**:

Verbesserung der betrieblichen Resilienz

Betriebliche Resilienz ist unerlässlich, damit Systeme reibungslos laufen. Dank der Unterstützung durch generative KI können IT-Teams die Ursachenanalyse beschleunigen und mehr Daten über alle Umgebungen hinweg korrelieren, um Probleme schneller zu erkennen. Außerdem steht ein spezielles Erkennungstool zur Beschleunigung der Reaktionen zur Verfügung – alles für den Erhalt der Geschäftskontinuität.

Stärkung der Kundenerlebnisse

Kundenzufriedenheit steht im Zentrum aller Unternehmen. Generative KI gibt Ihren Teams die Tools an die Hand, um Probleme schneller zu lösen und die Informationen zu erhalten, die sie brauchen. Gleichzeitig können sie Ihren Kunden persönliche Aufmerksamkeit widmen und schnellen Zugriff auf relevante Informationen bieten. Das Ergebnis? Verbesserte Kundenerlebnisse und bessere Geschäftsergebnisse.

Verringerung der Sicherheitsrisiken

Während sich die digitale Welt in atemberaubendem Tempo weiterentwickelt, entstehen neue und raffinierte Sicherheitsbedrohungen. Um ihnen entgegenzutreten, sind dynamische und proaktive Maßnahmen erforderlich, ganz zu schweigen von der Expertise, wie auf die Bedrohungen zu reagieren ist und sie bewältigt werden können. Generative KI kann nicht nur Ihr Sicherheitsteam und den operativen Betrieb stärken, sondern auch Warnmeldungen automatisieren und ein proaktives Vorgehen ermöglichen.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Generative KI kann in allen Brachen die vorhandenen Mitarbeiter- und Kundenerlebnisse durch die Bereitstellung personalisierter, relevanter und präskriptiver Antworten auf ihre Anfragen verbessern. Ganz egal, in welchen Sektor Sie tätig sind, es gibt immer eine Möglichkeit zur Operationalisierung von generativer KI, um eine leistungsstärkere Suche zu ermöglichen und Ihre Daten zur Erschließung neuer Möglichkeiten zu nutzen.

Telekommunikation

Für Telekommunikationsunternehmen wird prognostiziert, dass generative KI einen Wirtschaftswert von über 60 Milliarden Dollar schaffen wird.[4] Mithilfe von generativer KI können die Beschäftigten und Kunden von Telekommunikationsunternehmen ihre Website oder interne Wissensdatenbank abfragen, um schnell personalisierte und relevante Antworten zu erhalten. Das Ergebnis? Besserer Kundenservice und gesteigerte Produktivität.

[4] McKinsey, Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom, (2024).

Kundenerlebnis
Steigerung von Umsatz und Rentabilität

Optimiertes Sucherlebnis	Produkt-empfehlungen auf der Basis individueller Abfragen	→ Kundenerlebnis verbessern → Zufriedenheit steigern → Transaktionen pro Kunde erhöhen
Automatisierter Kundenservice	Self-Service-Support rund um die Uhr mittels Chatbots	→ Servicequalität insgesamt verbessern → Reaktionszeit bei Kundenanliegen verringern → Zufriedenheit und Kundenbindung steigern

Mitarbeitererlebnis
Produktivitätssteigerung und Kostenreduktion

Assistent der Wissens-datenbank	Abruf und Zusammenfassung von Informationen auf der Basis von generativer KI	→ Entscheidungsfindung beschleunigen → Zeitaufwand für manuelle Aufgaben reduzieren → Schnelle Synthese und Extraktion von Informationen
KI-Assistent für Netzwerke	Proaktive Empfehlungen und Abhilfemaßnahmen für Netzwerk-probleme	→ Betriebliche Effizienz durch Empfehlungen der generativen KI verbessern → Netzwerkausfallzeiten reduzieren → Kosten für Notfall-reparaturen reduzieren

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Finanzdienstleistungen

Mit generativer KI können Finanzdienstleister ihre Kunden- und Mitarbeitererlebnisse besser personalisieren. Es wird geschätzt, dass Verbesserungen bei Kundenerlebnis, Betrugsprävention und Automatisierung einen Wirtschaftswert von mehr als 250 Milliarden USD für die Finanzdienstleistungsbranche schaffen werden.[5]

[5]McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Kundenerlebnis
Steigerung von Umsatz und Rentabilität

Assistent für das Privatkundengeschäft	Abruf und Zusammenfassung von Informationen auf der Basis von generativer KI	→ Mehr Transparenz über persönliche Finanzen → Maßgeschneiderte Angebote für höhere Konversionsraten → Zufriedenheit und Kundenbindung steigern
Verbesserter Kundenservice	Proaktive Empfehlungen und Abhilfemaßnahmen für Netzwerkprobleme	→ Servicequalität insgesamt verbessern → Reaktionszeit bei Kundenanliegen verringern → Kundenbindung erhöhen

Mitarbeitererlebnis
Produktivitätssteigerung und Kosten- und Risikoreduktion

Zusammenfassung der Betrugserkennung	Anomalie erkennung/ Transaktionszusammenfassung und nächstbeste Aktion	→ Genauigkeit und Geschwindigkeit der Betrugserkennung verbessern → Kosten durch Automatisierung von Aufgaben reduzieren → Finanzielle Verluste reduzieren
Virtueller Assistent	Abruf und Zusammenfassung von Informationen auf der Basis von NLP	→ Entscheidungsfindung beschleunigen → Zeitaufwand für manuelle Aufgaben reduzieren → Schnelle Synthese und Extraktion von Informationen

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Einzelhandel

Für den Einzelhandel ist die generative KI besonders attraktiv, da sie eine höhere Kundenbindung verspricht, indem sie die Relevanz von Suchanfragen steigert, zusätzliche Produkte empfiehlt und personalisierte Follow-up-Nachrichten über alle Kanäle hinweg versendet. Haben Sie jemals eine dieser Mails mit folgendem Inhalt erhalten: „Sie haben etwas in Ihrem Einkaufswagen vergessen!“? KI kann solche Mails automatisieren und so verbessern, dass sie bessere Empfehlungen und mehr personalisierte Produktentdeckungen enthalten.

Ob es darum geht, Kundenerlebnisse der nächsten Generation zu schaffen, um den E-Commerce-Umsatz anzukurbeln, oder Mitarbeiter mit der neuesten Technologie auszustatten, um die Produktivität zu steigern: Laut den Prognosen wird generative KI einen Wirtschaftswert von über 240 Milliarden Dollar für Einzelhändler schaffen.[5]

[5]McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Kundenerlebnis
Steigerung von Umsatz und Rentabilität

Personalisierte Produktsuche und -entdeckung	Beantwortung von Fragen, maßgeschneidertes Sucherlebnis	→ Website-Konversionsraten verbessern → Transaktionen pro Kunde erhöhen → Zufriedenheit steigern
Verbesserter Kundenservice	Self-Service-Interaktionen mittels Chatbots	→ Reaktionszeit bei Kundenanliegen verringern → Besserer Service, weniger Abwanderung → Kundenbindung erhöhen

Mitarbeitererlebnis
Produktivitätssteigerung und Kostenreduktion

Verbesserter Kundenservice	Verbesserte Mitarbeitererfahrung und -interaktion	→ Problembehebung beim ersten Kontakt → Schnelleres Onboarding → Verringerung der Mitarbeiterfluktuation
Prädiktive Wartung	Bewertung des Zustands kritischer Systeme, um Prioritäten für wichtige Wartungsaufgaben zu setzen	→ Weniger Ausfallzeiten von Geräten und Systemen → Geringere Kosten für Notfallreparaturen → Verbesserte Betriebseffizienz

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Fallstudie: HSE

HSE ist eine der führenden Marken im europäischen Live-Commerce-Sektor.[6]

„Für Home Shopping Europe (HSE) beginnt der kommerzielle Erfolg mit der Personalisierung und Relevanz der Website."

Peter Strasser
Software Developer bei HSE

Die Chance

Wie bei jedem E-Commerce-Unternehmen ist die Suchfunktion für das Kundenerlebnis und den Umsatz von entscheidender Bedeutung. HSE muss dabei Customer Journeys aus den verschiedensten Kanälen berücksichtigen, die zu unterschiedlichsten Suchbegriffen führen, je nachdem, wo die Kunden das Produkt kennengelernt haben.

HSE verwendete generative KI und LLMs, um die semantische Bedeutung einer Kundenanfrage zu extrahieren und Ergebnisse zu generieren, die den herkömmlichen Abgleich von Suchbegriffen ergänzen.

Das Ergebnis

Dank eine höheren Genauigkeit und Relevanz der Suchergebnisse konnte HSE einen **4-prozentigen Anstieg bei der Click-Through-Rate** und eine **8-prozentigen Anstieg bei der Kundenzufriedenheit** erzielen.

Insight

Fokussieren Sie sich auf einen Bereich, für den Sie bereits Verbesserungen angestrebt haben, wie beispielsweise das Sucherlebnis der Kunden. Erfahren Sie, wie Sie generative KI integrieren können, um das Erlebnis durch Personalisierung und Relevanz auf die nächste Stufe zu heben.

[6] Elastic, HSE erhöht mit Elasticsearch auf AWS die Kundenzufriedenheit und reduziert die Wartungszeit um 42 %, (2024).

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Automobil- und Fertigungsbranche

Jeder Prozessschritt in der Automobil- und Fertigungsindustrie kann mit KI rationalisiert werden, was einen prognostizierten Wirtschaftswert von mehr als 170 Milliarden Dollar ergibt.[5] Generative KI hat das Potenzial, die Branche von der Produktforschung und Entwicklungsinnovation bis hin zu personalisierten Kundenbindungsstrategien zu verändern.
Fliegende Autos? Vielleicht!

[5]McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

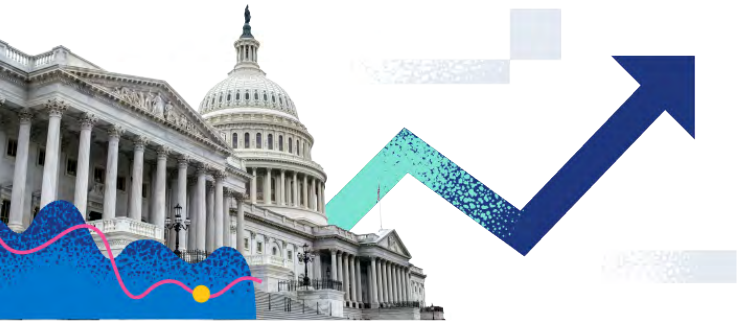
Kundenerlebnis
Steigerung von Umsatz und Rentabilität

Interaktive digitale Handbücher	Virtueller Produkt-Assistent	→ Echtzeitantworten zu Produkt-Features, Wartung und Fehlerbehebung → Support-Anfragen reduzieren → Zufriedenheit verbessern
Verbesserter Kundenservice	Self-Service-Interaktionen mittels Chatbots	→ Reaktionszeit bei Kundenanliegen verringern → Besserer Service, weniger Abwanderung → Kundenbindung erhöhen

Mitarbeitererlebnis
Produktivitätssteigerung und Kostenreduktion

Optimierung der Betriebs-technologie	Prädiktive Wartung: Zusammenfassung von Problem und Lösung	→ Probleme schnell identifizieren und beheben → Betriebseffizienz und Entscheidungsfindung verbessern → Fertigungskosten reduzieren
Sentimen-tanalyse für Produkte	Zusammenfassung des Produktanteils und Empfehlen von Verbesserungen	→ Produktangebot mit Kunden-PoV verbessern → Time-to-Value für neue Produktangebote verringern

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.



Öffentlicher Sektor

Generative KI kann Missionsergebnisse erheblich beschleunigen, Bürgerdienstleistungen verbessern und für Behörden tätigen Analysten und Security-Fachkräften die richtigen Daten zur richtigen Zeit an die Hand geben, indem generative KI auf sichere Weise mit den Behördendaten kombiniert wird.

Anwendungen für Bürgerdienste beinhalten:

- Personalisierter Zugriff auf öffentliche Dienstleistungen
- Vereinfachte Online-Erfahrungen für Bürger
- Bessere Zugänglichkeit und Self-Service-Optionen

Mitarbeiterorientierte Anwendungen beinhalten:

- Genauere Untersuchungen und Informationen
- Verbesserte Produktivität durch die Automatisierung manueller Prozessen und Workflows
- Effizientere Beschaffungsprozesse

Reduzierung der Arbeitslast
Automatisierung manueller Prozesse und Workflows

Compliance
Aktivierung von rollenbasiertem Datenzugriff

Echtzeit-Situationsbewusstsein
Eine genauere Datengrundlage für Entscheidungen

Mitarbeiterproduktivität
Die richtigen Informationen zur richtigen Zeit

Bürgererfahrungen
Vertrauensaufbau über personalisierte digitale Interaktionen

Öffentliche Dienstleistungen
Bessere Zugänglichkeit und Self-Service-Optionen

Dynamische Intelligenz
Beschleunigung der Suche und Insights für Ihre Mission

Cybersicherheit
Durchführung einer Risikobewertung und -analyse in Echtzeit

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Fallstudie: Relativity

Relativity hilft Unternehmen, Anwaltskanzleien und Behörden bei der Speicherung und Nutzung von Daten für e-Discovery und juristische Recherchen.[7]

„Die größte Herausforderung, der die Kunden von Relativity aktuell gegenüberstehen, ist die exorbitante Zunahme an Daten aus heterogenen Datenquellen. Die Herausforderung wird durch die Unterschiede in den Daten, die auf die verschiedenen Kommunikationsmittel zurückzuführen sind, noch verschärft."

Brittany Roush
Senior Product Manager

Die Chance

Relativity musste seine Daten konsolidieren, dabei aber der Sicherheit oberste Priorität einräumen. Infolge dieser exorbitanten Zunahme an Daten, Quellen und Komplexität waren herkömmliche Methoden der Schlüsselwortsuche ineffektiv. Zeit für RAG.

Das Ergebnis

Zusammen mit RAG und einer Vektordatenbank hat Relativity Sucherlebnisse implementiert, die auf proprietären Daten aufbauen und den Nutzern ein schnelles, relevantes und genaues Sucherlebnis bieten. Seine Lösung für generative KI erfüllt die Compliance-Standards wie PCI, DSS, SOC2 und HIPAA.

Insight

Behalten Sie bei der Konzeption künftige Skalierungen im Blick. Es kann hilfreich sein, klein zu beginnen, um die Fähigkeiten der generativen KI kennenzulernen und sich auf die relevantesten Anwendungen zu konzentrieren. Sobald man die Vorteile kennt, gibt es keine Grenzen mehr.

[7] Elastic, Relativity nutzt Elasticsearch und Azure OpenAI zur Erstellung futuristischer Sucherlebnisse, heute (2024).

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Sie verstehen nun das enorme wirtschaftliche Potenzial, das die generative KI über alle Branchen hinweg bietet. Vielleicht haben Sie bereits mögliche Anwendungsfälle im Kopf. Außerdem sind Sie sich hoffentlich über das „Warum“ klar.

Die Implementierung von generativer KI kann dennoch als mühevoller und disruptiver Prozess erscheinen. Es gibt Datenschutzbedenken, einiges an Arbeit im Bereich Compliance und außerdem müssen die Beschäftigten ihre Art der Aufgabenausführung ändern. Eine verantwortungsvolle Operationalisierung benötigt Schulung, Fortbildung und eine teilweise Umorganisation der Belegschaft.

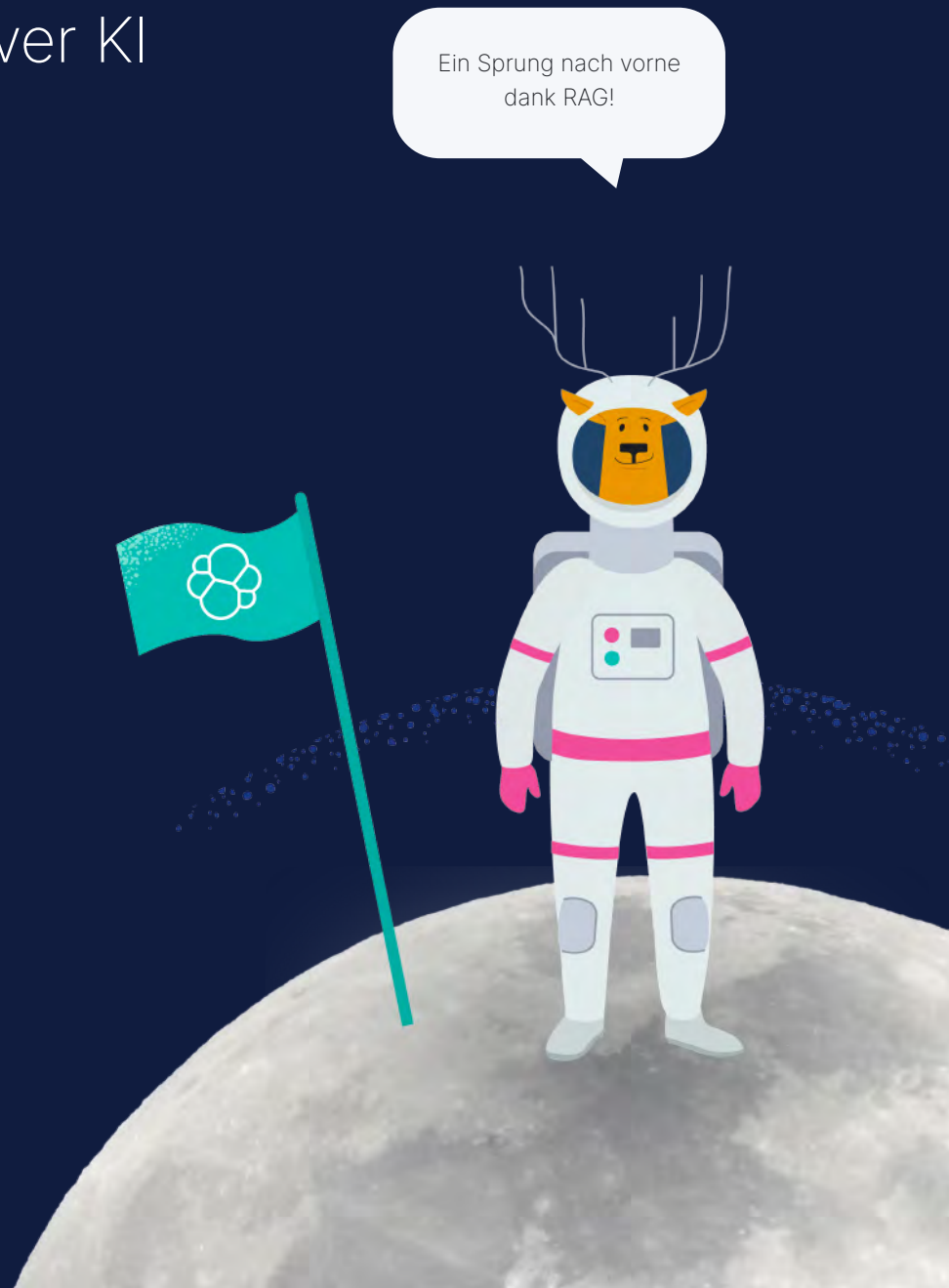
Trotz dieser Herausforderungen ist der Wert, den ein Unternehmen durch generative KI erzielen kann, unbestreitbar. Um wettbewerbsfähig zu bleiben, ist die Implementierung unverzichtbar. Die gute Nachricht? Sie können mit Tests, die noch nicht ganz produktionsreif sind, eine schnelle Wertschöpfung erzielen. Mit anderen Worten: Es ist Zeit für die ersten Schritte.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Teil 2:

Operationalisierung von generativer KI

Ein kleiner Schritt für die Maschine – ein großer Schritt für Ihr Unternehmen



Sie können generative KI nicht auf einen Schlag operationalisieren, sondern es ist ein schrittweiser Prozess, der Planung und ein klares Ergebnis erfordert. Indem Sie zu Beginn generative KI nur für ein einziges Projekt einsetzen – d. h. einen kleinen Schritt machen – geben Sie Ihrem Team die Möglichkeit, die unvermeidliche Lernkurve zu durchlaufen und die Prozesse zu erproben, die Technologie zu verfeinern und Bedenken auszuräumen. Auf diese Weise werden die Prozesse und ihr Unternehmen auf den Erfolg vorbereitet: den begehrten großen Sprung.

Wie das geht, erfahren Sie hier.

elastic.co/de | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Schritt 1 Identifizieren Sie Ihr ideales Ergebnis

Sie haben ein Problem erkannt. Sie wissen, dass Sie einen ineffizienten Prozess optimieren müssen. Sie müssen nun darüber nachdenken, wie die Interaktion der Benutzer mit Ihrer Lösung aussehen wird. Erweitern Sie eine Suchanwendung oder einen Chatbot? Sind Sie auf der Suche nach einer neuen Möglichkeit der Interaktion mit Ihren Teams oder Kunden?

Ihr Gedankengang sieht möglicherweise so aus:

- Sie möchten eine höhere Kundenbindung.
- Sie haben beschlossen, eine personalisierte Produktsuche und Discovery-Anwendung zu implementieren.
- Sie erstellen Erfolgsmetriken. Stellen Sie sich dies als Ihr „untergeordnetes Warum“ vor.

  Sie möchten generative KI verwenden. Warum? Zur Personalisierung der Produktsuche und -entdeckung. Warum? **Hier ist Ihr ideales Ergebnis:** Indem Sie auf diese neue Art mit unseren Daten interagieren, finden Kunden mühelos die von ihnen benötigten Produkte und entdecken in Abhängigkeit von ihrem Suchverlauf und Standort Produkte, die für sie interessant sein könnten. Infolgedessen wird die Kundenbindung verbessert.
- Nun starten Sie die große Aufgabe der Operationalisierung Ihres ersten Projekts mit generativer KI.

Stellen Sie sich folgende Frage:

Welche Aktionen und Ergebnisse können aus dieser neuen Interaktionsart mit Ihren Daten entstehen?

Die Antwort auf diese Frage hilft Ihnen bei der Festlegung von Zielvorgaben. Durch die Identifizierung des idealen Ergebnisses bestimmen Sie, was „gut“ für Ihr Projekt und – in größerem Maßstab – für Ihr Unternehmen bedeutet.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Schritt 2

Ermitteln Sie die Auswirkungen. Messen Sie den Erfolg.

Um bei der Operationalisierung von generativer KI erfolgreich zu sein, müssen Sie eine Reihe von KPIs festlegen, mit deren Hilfe Sie messen können, was „gut“ für Sie bedeutet. Zu verstehen, wie generative KI die Produktivität in Ihrem Unternehmen nach oben treibt, ist nur eine der wichtigsten Leistungskennzahlen.

Andere Beispiele sind die Steigerung der Kundenzufriedenheit, die anhand von Bewertungen im Rahmen des Kundensupports gemessen wird, ein Rückgang der Support-Tickets oder eine schnellere Problemlösung. Je nachdem, welchen Anwendungsfall Sie erproben, müssen Sie die zugehörigen Leistungsindikatoren festlegen. Die Einbettung dieser Kennzahlen in jeden Schritt des Erprobungsprozesses ist unerlässlich, um die Fortschritte nachzuvollziehen, die Sie und Ihr Team machen.

Grundlegende Leistungsindikatoren

1

Auswirkungen auf die Produktivität

Messen Sie die Änderungen bei der Produktivität, die sich aus Ihrem Anwendungsfall ergeben. Vergleichen Sie die Zeit, die Sie zur Ausführung bestimmter Aufgaben benötigen, mit der Zeit, die Sie ohne die Nutzung generativer KI aufwenden.

2

Skalierbarkeit

Bewerten Sie, wie gut sich das Modell bei einem Anstieg der Nutzung und der Nachfrage skalieren lässt. Liefert es weiterhin zuverlässige und genaue Ergebnisse?

3

Fazit

Bewerten Sie, welche Auswirkungen die Implementierung von generativer KI im Hinblick auf die Unternehmenskosten hatte. Bei dieser Bewertung möchten Sie vielleicht bestimmte Unternehmenskennzahlen heranziehen, wie beispielsweise die Zahl der protokollierten Kundenbeschwerden oder Änderungen bei Ihren Umsätzen.

4

Compliance

Überwachen Sie kontinuierlich, ob beim Einsatz der generativen KI die Datenschutzbestimmungen eingehalten werden.

5

Kundenzufriedenheit

Prüfen Sie Unternehmenskennzahlen wie Kundenabwanderung, Umsatzsteigerungen und Aufrechterhaltung der Markentreue. Untersuchen Sie auch das Feedback der Kunden.

Verwenden Sie diese Kennzahlen, um festzustellen, ob ein Projekt machbar, durchführbar, skalierbar und bezahlbar ist. Diese Kennzahlen helfen Ihnen, Ihren ROI zu ermitteln, und können mit der Erweiterung Ihrer Anwendungsfälle in der Zukunft breiter gefasst werden.

Schritt 3 Suchen Sie ein Modell aus (Weg nach vorne)

Wie erstellen Sie aus generativer KI eine Architektur, die Ihre Geschäftsanforderungen erfüllt? Viele Faktoren fließen in Ihre Entscheidung ein: die Kosten, die Sprache, Ihr IT-Ökosystem, Ihre Deployment-Fähigkeiten und die Zeitleiste, Datenschutzbestimmungen und Governance. Aus diesem Grund ist es wichtig, sich langsam heranzutasten: Beginnen Sie mit einem einfachen und konkreten Anwendungsfall.

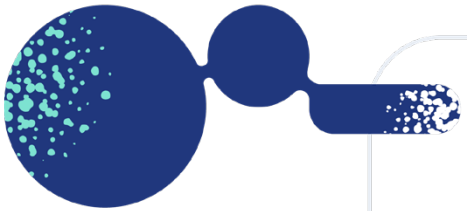
Um generative KI zu operationalisieren, benötigen Sie die folgenden Komponenten:

- **Eine vollständig verwaltete Cloud-Infrastruktur** erhöht die Agilität, verbessert die Kosteneffizienz und verringert den Ressourcenverbrauch. Chips und Hardware entwickeln sich in atemberaubendem Tempo weiter. Wenn Sie in den Bau eines eigenen KI-Rechenzentrums investieren, wird es möglicherweise bereits nach wenigen Monaten veraltet sein.
- **Ein LLM** bildet die Grundlage, die es der generativen KI ermöglicht, in natürlicher Sprache zu kommunizieren und sie zu verstehen.
- **Eine Datenplattform**, die Vektor-, Hybrid- und herkömmliche Schlüsselwortsuche bietet und verwendet werden kann, um dem LLM Ihre proprietären Daten als richtigen Kontext bereitzustellen.
- **Umfangreiche APIs**, mit deren Hilfe Sie Ihre Daten erweitern und an das LLM und Ihre Suchmaschine übertragen können.

Die Zutaten, die Unternehmen für die KI-Suche benötigen

Die Art und Weise, wie Sie diese Komponenten kombinieren – ob Sie Ihr Modell feintunen, Ihre eigene Vektordatenbank oder Ihr eigenes Modell mitbringen bzw. eine beliebige Kombination daraus – all das hat Auswirkungen auf Ihr Implementierungsschema, Ihre Zeitleiste, die Komplexität Ihrer Erprobung und erfordert möglicherweise eine Aufstockung Ihres Teams.

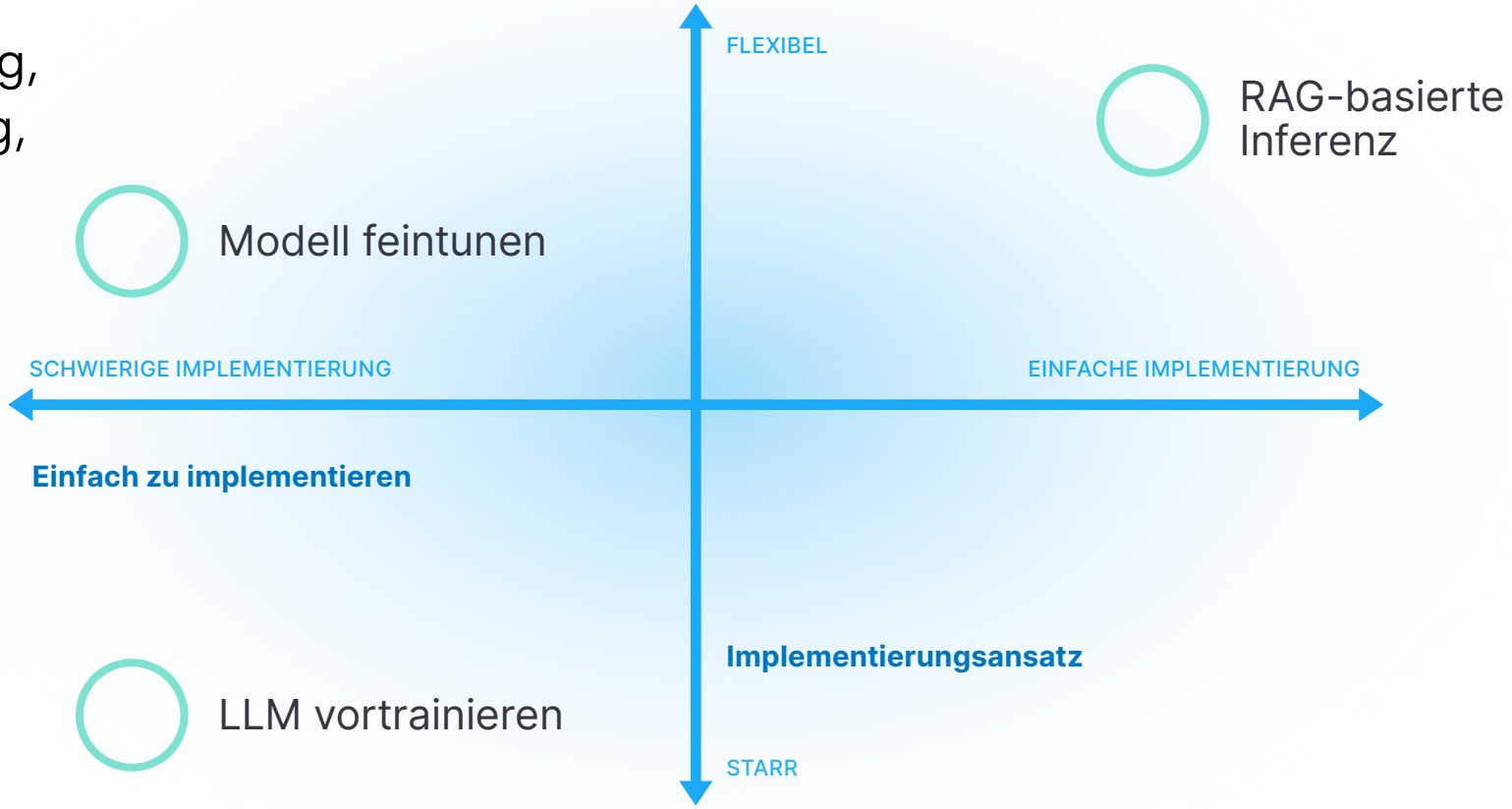
 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.



	Vortraining eines LLMs	Feintuning eines Modells	RAG
	Bei diesem ressourcenintensiven Ansatz beginnen Sie bei Null an, indem Sie ein großes Sprachmodell mit einem großen Datensatz trainieren.	Dabei wird ein vorhandenes LLM mit Ihrer Suchmaschine und einer Vektordatenbank verwendet, um Ihren proprietären Daten Kontext hinzuzufügen.	Dieser Prozess nutzt ein vorhandenes und vortrainiertes LLM sowie verschiedene Techniken, um das Modell auf Ihre Anforderungen abzustimmen.
Kosten	€€€€	€€€	€€
Bereitstel-lungszeit	Lange, mehrere Monate	Mittelfristig, innerhalb von Wochen	Schnell, innerhalb von Tagen
Sicherheit und Schutz Ihrer Daten	Ist Ihr Datensatz groß genug, um dem LLM signifikantes Lernmaterial bereitzustellen? Wenn nicht, brauchen Sie öffentliche Daten. Möchten Sie öffentliche und private Daten kombinieren?	Ist Ihr Datensatz groß genug, um dem LLM signifikantes Lernmaterial bereitzustellen? Wenn nicht, brauchen Sie öffentliche Daten. Möchten Sie öffentliche und private Daten kombinieren?	Dieser Ansatz ermöglicht Ihnen, Ihre privaten Daten privat zu halten.
Genauigkeit und Relevanz	Schwierig, dies ständig sicherzustellen	Es ist einfacher, Genauigkeit und Relevanz für die konkreten Aufgaben sicherzustellen, auf die das Modell abgestimmt wurde.	Der Vorteil von RAG beruht insbesondere auf der Fähigkeit, Halluzinationen zu reduzieren, indem es Quellen zitiert oder die Nutzer wissen lässt, wenn es keine Antwort hat.

Erwägen Sie diese Optionen:

Vortraining, Feintuning, RAG



 © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Auswahl des richtigen Wegs nach vorne

Es gibt nicht einen richtigen Weg für alle, daher müssen Sie sicherstellen, dass Sie Ihre Entscheidungen anhand der in früheren Schritten festgelegten Ziele prüfen und dass diese Ziele mit den betroffenen Parteien abgestimmt sind. Letztendlich brauchen Sie einen Weg, den Sie den Stakeholdern Ihres Unternehmens und Ihrem Team klar aufzeigen können.

Wenn Sie weder eine umfangreiche Überarbeitung noch ein sehr großes Projekt planen, ist es zu ressourcenintensiv, ein LLM von Grund auf neu aufzubauen und feinzutunen. Viele Fragen werden auf Sie einstürmen: Benötigen Sie eine Vektordatenbank zur Ergänzung Ihrer Suchmaschine? Können Sie Ihre Suchmaschine aufrüsten, in ihr Einbettungen erstellen und speichern und eine Logik entwickeln, um Ihre Suche weiterhin zu unterstützen? Wie können Sie dies zu einem Empfehlungssystem ausbauen?

Für cloudbasierte Lösungen mit hybrider und semantischer Suche ist das ziemlich einfach: Indem Sie eine Verbindung zu einem vorhandenen LLM herstellen, können Sie RAG nutzen, um Ihren Kunden ein relevanteres Sucherlebnis zu bieten.

Gehen Sie folgendermaßen vor:

- Machen Sie eine Bestandsaufnahme von dem, was Sie bereits in Ihrer IT-Umgebung haben. Häufig ist es gar nicht nötig, die Architektur der Infrastruktur neu zu erstellen.

- Überlegen Sie, ob Out-of-the-Box-Lösungen (OOTB) wie ein OOTB-LLM, eine OOTB-Vektordatenbank und ein OOTB-Komplettpaket in Frage kommen.

  Pro: Black-Box-Technologien lassen sich schneller einrichten und in Betrieb nehmen.

  Kontra: Sie lassen sich nicht gut skalieren, weil sie nur wenig individuelle Einstellmöglichkeiten bieten.

- Suchen Sie nach einem ergänzenden Produkt, das Flexibilität bei minimaler Betriebsunterbrechung bietet. Sie sollten Ihre Suchrelevanz und -leistung vergleichen und vielleicht die Modelle austauschen, um herauszufinden, welches für Sie am besten funktioniert.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Schritt 4

Schnelle Erprobung, schnelles Scheitern

Ein sich schnell entwickelndes digitales Ökosystem bedeutet, dass es bei einem Projekt mit generativer KI viele bewegliche Faktoren gibt. Die Möglichkeiten der Kontrolle, die Sie über ein vortrainiertes LLM haben, sind begrenzt, ebenso wie die Flexibilität bei der Manipulation Ihrer Architektur.

An diesem Punkt sollten Sie ein iteratives Herangehen erwägen: Sie haben Ihren Anwendungsfall, Sie haben die gewünschten Ergebnisse festgelegt, Sie haben KPIs bestimmt und Sie haben überlegt, wie Sie Ihr Projekt für generative KI umsetzen wollen.

Denken Sie daran:

Im Grunde geht es bei der Operationalisierung von generativer KI darum, Antworten aus *Ihren* Daten zu erhalten. Behalten Sie die Compliance stets im Blick. Könnte der Test, den Sie gerade einrichten, Ihre Datenschutzrichtlinien verletzen? Ist dies ein Test mit niedrigem Risiko?

Zu diesem Zeitpunkt möchten Sie ...

- **Eine Feedback-Schleife erstellen:** Legen Sie fest, wer was an wen berichtet, und ermitteln Sie die wichtigsten Stakeholder in dem Projekt.
- **Ihr LLM anreichern:** Stellen Sie sicher, dass Ihr LLM Zugriff auf die richtigen Informationen hat, die in einer Vektordatenbank gespeichert werden. Mit einer Vektordatenbank können Sie schnell auf die relevantesten Informationen zugreifen, um damit Ihr LLM anzureichern.
- **Benutzererlebnis feintunen:** Arbeiten Sie an einer benutzerfreundlichen Oberfläche und führen Sie immer wieder Tests durch. Letztendlich ist generative KI dazu da, um Ihren Mitarbeitern und Kunden das Leben zu erleichtern. Die Gestaltung einer Oberfläche, die zur Anwendung und den Nutzern passt, ist für einen erfolgreichen Einsatz von generativer KI von entscheidender Bedeutung und stellt die Skalierbarkeit sicher.
- **Eine skalierbare Referenzarchitektur einrichten:** Während Sie den Einsatz der generativen KI testen, müssen Sie das Gesamtbild im Auge behalten. Wie wird Ihre Architektur aussehen, wenn Sie das Projekt skalieren, und wie wird sie aussehen, wenn Sie weitere Anwendungsfälle erproben?

Wenn Sie beispielsweise die Mühe scheuen, eine Vektordatenbank von Grund auf zu erstellen, können Sie eine downloadbare Version erwägen – so etwas gibt es tatsächlich. Mit einer solchen Vektordatenbank erreichen Sie die nächste Ebene: die Hybridsuche. Indem Sie bei Ihren Suchanwendungen dies semantische Suche verwenden, können Sie den Prototypen für Ihr KI-Projekt der nächsten Generation testen. Dies ist ein Beispiel dafür, wie wichtig es ist, klein anzufangen und zu den Prozess immer wieder zu wiederholen.

 © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Schritt 5 Governance und operativer Betrieb

Mit dem Einsatz generativer KI gehen ganz besondere Herausforderungen einher – von Datenschutz und Compliance über ethische Erwägungen bis hin zu Qualitätskontrolle und Risikomanagement. Sie müssen mögliche Hindernisse voraussehen und sicherstellen, dass Ihr Projekt an Ihren Unternehmenszielen ausgerichtet ist.

Bei der Prüfung Ihrer Governance und Ihres operativen Betriebs müssen Sie verschiedenste Elemente berücksichtigen:

- **Kostenverwaltung:** Die Abrechnung erfolgt pro Tausend Token; eine Gebühr für die Eingabeaufforderung, eine für die Antwort.
- **Logging:** Jede Antwort muss geloggt werden, um zur Qualitätskontrolle die Kommunikation zwischen Ihrem Modell und Ihrem Kunden nachverfolgen zu können.
- **Die Stimmung der Antworten festlegen:** Legen Sie fest, welche Stimmung die LLM-Antworten vermitteln sollen, damit sie mit dem Tonfall Ihres Unternehmens übereinstimmen (ein weiterer wichtiger Schritt zur Qualitätskontrolle).
- **Auf Halluzinationen achten:** Halluzinationen beinhalten falsche oder irreführende Informationen, können jedoch auch Hasssprache und antisoziales Verhalten eines Chatbots umfassen.
- **Unschlüssige Antworten markieren:** Die Überwachung der Qualität und Relevanz der Antworten ist für die Qualitätskontrolle entscheidend. Dies bietet die Möglichkeit, zu verstehen, welche Anwendungen mehr menschliche Beteiligung erfordern als andere, und entsprechend zu planen, wenn eine Skalierung ansteht.

Voreingenommenheit in der KI

Modelle der generativen KI basieren auf den Daten, mit denen sie trainiert wurden. Wenn die Trainingsdaten voreingenommen oder eingeschränkt sind, spiegelt sich dies in den Outputs wieder.

Unternehmen können diesen Risiken begegnen, indem sie die Trainingsdaten für ihre Modelle sorgfältig auswählen und einschränken oder indem sie maßgeschneiderte und spezialisierte Modelle für Ihre Anforderungen einsetzen. Allerdings sind auch die Menschen, die diese Technologie programmieren oder die Daten aufbereiten, mit denen das Modell trainiert wird, voreingenommen.

Voreingenommenheit lässt sich nur schwer ausschließen, dies gilt für jeden Kontext. Das heißt nicht, dass Unternehmen sich dieser Herausforderung nicht stellen und die Nutzer dahingehend schulen sollten, dass sie im Rahmen der Lösung kritisches Denken walten lassen.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Gehen Sie außerdem von einer Einbeziehung Ihres Rechtsteams aus, und stellen Sie sicher, dass Sie dessen Arbeit in Ihrem Machbarkeitsnachweis berücksichtigen. Obwohl die Einbeziehung dieses Teams den Eindruck erwecken mag, die Testphase zu verlangsamen, ist sie dennoch entscheidend, um Prüfungsprozesse zu etablieren, die gründlich und effizient für eine verantwortungsvolle, ethische und regelkonforme Implementierung sind.

Die Sache mit der Datensicherheit

Angesichts der Sicherheitsbedrohungen, mit denen Unternehmen tagtäglich konfrontiert sind, ist Datensicherheit unerlässlich. Ihre Kunden vertrauen Ihnen ihre Daten an. Aus diesem Grund entscheiden sich viele Unternehmen für ein Zero-Trust-Framework. Dieses beruht auf dem Grundsatz, dass Nutzer und Geräte niemals automatisch oder implizit vertrauenswürdig sind, egal ob sie sich innerhalb oder außerhalb der Netzwerkparameter eines Unternehmens befinden.

Zur Optimierung Ihrer Sicherheit haben Sie folgende Möglichkeiten:

1. **Verwenden Sie einen RAG-Ansatz:** RAG-Modelle nutzen Abrufmechanismen, um den Kontext der Eingabeaufforderung besser zu verstehen und so kontextgerechtere Antworten zu liefern, die keine sensiblen Daten enthalten. Die Verwendung von RAG mit einer Datenplattform, die Security auf Dokumentenebene und Rollenbasis bietet, stellt sicher, dass Berechtigungen eingehalten werden.

2. **Investieren Sie in Ihre Observability-Lösung oder in ihre Erweiterung:** Stellen Sie sich die Frage nach dem Vertrauen. Folgen Sie dem Datenpfad mit Überwachungsfunktionen und überwachen Sie die Antworten, die Ihre generative KI erstellt hat. Wohin gehen Ihre Daten und was sagt die generative KI zu Ihren Kunden?

Um generative KI in Ihrem Ökosystem einzuführen, müssen Sie neue Betriebsprotokolle und entsprechend auch neue Richtlinien einführen. Mit effizienteren Prozessen und höheren Erträgen können Sie die Zeit, die Sie beim Ausführen alltäglicher Aufgaben sparen, in diese Bemühungen investieren.

Schritt 6 Legen Sie eine Zeitleiste fest. Geben Sie Benchmarks vor.

Legen Sie eine Zeitleiste fest, beispielsweise ein Vierteljahr. Legen Sie innerhalb dieser Zeitleiste Zielvorgaben für den 30. und den 90. Tag fest. Nutzen Sie das Vierteljahr, um nachzuweisen, welchen Wert Ihr Anwendungsfall für generative KI erzielt.

Bis zum 30. Tag sollten Sie Ihren ersten Test gestartet haben. Wie sieht es aus?

- [ ] Sie haben einen Anwendungsfall ausgewählt
- [ ] Sie haben der Aufgabe ein kleines Team zugewiesen
- [ ] Sie haben Schulungssitzungen nach Bedarf ermöglicht
- [ ] Sie haben die gewünschten Ergebnisse festgelegt
- [ ] Sie haben eine Prototyp-Oberfläche erstellt

Bis zum 90. Tag sind Sie bereit, Ihren ersten Anwendungsfall zu starten. Wie sieht es aus?

- [ ] Sie haben den Test für einige interne Mitarbeiter geöffnet
- [ ] Sie haben die erzeugten Outputs getestet, abgestimmt und gemessen
- [ ] Sie haben kontinuierlich überwacht, wie Nutzer mit der Oberfläche interagieren
- [ ] Sie haben in einer Reihe von Richtlinien festgelegt, was ein Qualitäts-Output ist
- [ ] Sie haben Daten zu wichtigen Leistungskennzahlen erfasst
- [ ] Sie haben den Wert der Initiative gemessen

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Diese Aufgaben sollten als grobe Benchmarks dienen. Die speziellen Anforderungen Ihres Unternehmens – die Zusammensetzung Ihres Teams und die Technologie, an der es arbeitet oder die es Ihrem Stack hinzufügt – werden sich auf die Geschwindigkeit auswirken, mit der Sie Ihren ersten Anwendungsfall implementieren und Erkenntnisse gewinnen können.

Betrachten Sie an diesem Punkt Folgendes:

1. **Die Fehlerquote:** Messen Sie die Fehlerquote. Produziert die generative KI richtige und relevante Outputs? Dies ist für das Feintuning der generativen KI von entscheidender Bedeutung.
2. **Trainingszeit und -kosten:** Messen Sie die Zeit und Ressourcen, die Sie zum Trainieren Ihres Modells benötigen. Dies trägt zu einer effizienten Testphase und damit zu einer schnelleren Operationalisierung bei.
3. **Menschliche Eingriffe:** Können Sie die generative KI nur mit Human-in-the-Loop ausführen? Wie viel Überwachung ist benötigt, um Zuverlässigkeit und Genauigkeit aufrechtzuerhalten?
4. **Reaktionszeit und Qualität der Outputs:** Messen Sie, wie schnell die generative KI Outputs liefert und vergleichen Sie die Qualität der Outputs mit verschiedenen festgelegten Regeln oder Richtlinien.

Und schon sind Sie bereit für die Operationalisierung und Skalierung Ihres Erfolgs.

 © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Der Beginn einer neuen Ära

Viele Branchenführer können bereits von generativer KI profitieren und noch mehr bemühen sich gerade darum, diesen Erfolg zu wiederholen und mit den sich wandelnden Kundenerwartungen Schritt zu halten. Die Innovation der generativen KI schreitet schnell voran. Ohne Grundlagenkenntnisse können Sie jedoch gar nichts machen.

Wenn Sie eine Strategie für die am besten abgestimmten Implementierungsmöglichkeiten für generative KI entwickeln, können Sie die Macht Ihrer Daten nutzen, ohne sich von spannenden, aber irrelevanten Innovationen ablenken zu lassen. Für die effektivste Operationalisierung von generativer KI ist es am besten, Zeit und Ressourcen in eine schrittweise Implementierung zu investieren. Die Integration neuer zielgerichteter Technologien ist das Rezept, um aus Ihrer Investition die höchste Rendite zu erwirtschaften. Darüber hinaus sorgt die individuelle Anpassung von KI-Tools an Ihre betrieblichen Anforderungen für Relevanz und Effektivität – und genau das ist die eigentliche „Berufung“ der generativen KI.

Bedenken Sie jedoch stets die Notwendigkeit, generative KI auf verantwortungsvolle Weise zu implementieren – die gilt für die Sicherheit und den Schutz der Daten ebenso wie für Sensibilität und Ethik. Abgesehen davon, dass generative KI die Weltwirtschaft um Billionen von Dollar bereichert, eröffnet sie auch die Möglichkeit für eine Demokratisierung und Weiterbildung der Arbeitskräfte. Positionieren Sie sich nicht nur als Vorreiter für generative KI, sondern gehen Sie auch bei der Entwicklung neuer Geschäftsprozesse für Ihr Unternehmen voran.

 | © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

Lassen Sie uns anfangen.

Es ist eine teamübergreifende Aufgabe, Ihren ersten Anwendungsfall für generative KI auszuwählen. Sie müssen sicherstellen, dass Ihr Sicherheitsteam, Ihr IT-Team, Ihr Entwicklerteam und das Team des jeweiligen Geschäftsbereichs vom ersten Tag an zusammenarbeiten. **So kann Elastic helfen:**

Mit Ihrem Sicherheitsteam

Sie können die Produktivität der Praktiker steigern und das Risiko reduzieren. Die Operationalisierung von generativer KI für Ihre Security-Anwendungsfälle beginnt mit einer einheitlichen Herangehensweise auf einer offenen Plattform. Mit Elastic Security können Sie das Potenzial der generativen KI nutzen, um ein Erlebnis zu erstellen, das speziell auf die Bedürfnisse Ihres Sicherheitsteams abgestimmt ist.

Elastic Security entdecken

Mit Ihrem SRE- und IT-Operations-Team

Geben Sie Ihren SREs und Technikern die Möglichkeit an die Hand, eine Chat-Oberfläche mit interaktiver natürlicher Sprache zu nutzen, die ihnen einen schnelleren Zugriff auf die relevantesten Informationen bietet. Erfahren Sie, wie Sie Conversational AI mit Elastic Observability und innovativen Machine-Learning-Funktionen für ein kontextbezogenes interaktives Chat-Erlebnis auf der Basis Ihrer eigenen Daten und Runbooks kombinieren können.

Elastic Observability entdecken

Mit Ihrem Entwicklerteam

Sie können das Toolkit Ihres Entwicklerteams um Self-Service-Optionen für den Kundensupport erweitern, wie beispielsweise hoch personalisierte Chatbots mit Elastic Search. Geben Sie diese großartigen Suchtools auch Ihren Kundenservice-Mitarbeitern an die Hand, damit sie Fälle schneller lösen können. Dazu gehören auch Erlebnisse der generativen KI, die ihnen helfen, Antworten aus unterschiedlichen Datenquellen zu finden. Entdecken Sie, wie Sie eine leistungsstarke Suche für Ihre Wissensdatenbank implementieren können.

Elastic Search entdecken

 © 2024 Elasticsearch B.V. Alle Rechte vorbehalten.

elastic | The Search AI Company