# Mit MongoDB Atlas generative KI und erweiterte Suche in Ihre Apps einbetten

## KI-basierten Anwendungen entwickeln

Dezember 2023

US 866-237-8815 • INTL +1-650-440-4474 • info@mongodb.com.

# Inhaltsverzeichnis

# Einführung

Nie zuvor hat die Einführung einer neuen Technologie so schnell die Aufmerksamkeit von Unternehmen, Regierungen und Verbrauchern gleichermaßen auf sich gezogen. Die Einführung von ChatGPT im November 2022 hat das Potenzial von generativer KI auf der Grundlage von Large Language Models (LLMs) für eine Vielzahl neuer Anwendungsfälle aufgezeigt. Diese Anwendungsfälle waren zuvor mit herkömmlicher Computer- und analytischer KI (heute manchmal als „traditionelle“ oder „klassische“ KI bezeichnet) unvorstellbar.

Alles, was nötig zu sein scheint, sind ein paar gut ausgearbeitete Eingabeaufforderungen, um eine ganze Reihe von Dingen zu automatisieren. Generieren Sie professionellen Text, Bilder, Audio, Video und Programmiercode. Kunden besser unterstützen. Das reicht von der Modellierung des Klimawandels über die Entdeckung neuer Medikamente oder die Entwicklung neuer Materialien bis hin zur Vorhersage der Entwicklung der Finanzmärkte und viele andere Bereiche.

Über Nacht tauchte eine Frage ganz oben auf der Tagesordnung jeder Vorstandsetage auf: *„Wie können wir generative KI nutzen, um unsere Märkte zu revolutionieren, ohne dabei selbst abgehängt zu werden?“*

Technologieführer haben jedoch schnell erkannt, dass neben den potenziellen Vorteilen von generativer KI auch Risiken durch die Unausgereiftheit der Technologie vorhanden sind. Jahrelange betriebliche Best Practices und institutionelles Wissen können nicht einfach über Bord geworfen werden. Stattdessen muss sichergestellt werden, dass sowohl die bestehenden Systeme als auch die neu entwickelten Anwendungen in der Lage sind, die generative KI sicher, zuverlässig und präzise zu nutzen.

In diesem Beitrag erörtern wir, wie Sie mit MongoDB diese Ziele erreichen und gleichzeitig Ihre eigenen Daten nutzen können, um überzeugende neue KI-gestützte Anwendungen und Erlebnisse zu schaffen.

# Kontext ist alles

Wenn jeder Zugang zu generativen KI-Modellen hat, besteht Ihre „Superkraft“ darin, dass Sie diesen Modellen Zugang zu einem Ihrer wichtigsten Unternehmensressourcen geben – nämlich Ihren Daten. Einige dieser Daten sind Eigentum des Unternehmens, andere sind öffentlich – aber aktueller – als die Daten, die zum Trainieren der ursprünglichen Foundation-Modelle verwendet wurden. Zusammen liefern diese Daten Antworten, die die heutige Realität besser widerspiegeln.

Die Versorgung von Modellen mit Ihren eigenen Daten wird durch ein neues Architekturmuster namens Retrieval-Augmented Generation (RAG) erreicht. Der Einsatz von RAG bietet Ihren Entwicklern eine leistungsstarke Kombination. Sie können das unglaubliche Wissen und die Schlussfolgerungsfähigkeiten von vortrainierten, allgemeinen generativen KI-Modellen nutzen und sie mit genauen und aktuellen unternehmensspezifischen Daten füttern.

Das Ergebnis sind generative KI-Ausgaben, die genau, aktuell und relevant sind und alle Ihre Daten nutzen, unabhängig von ihrer Struktur. Ihre KI-gestützten Apps bieten einen besseren Service für Ihre Kunden, steigern die Produktivität Ihrer Mitarbeiter und sind innovativer als die Konkurrenz. Ihre Entwickler können all diese Ergebnisse nutzen, ohne sich an spezialisierte Data Science Teams wenden zu müssen, um Modelle zu trainieren oder zu optimieren – ein komplexer, zeitaufwändiger und teurer Prozess.

Die Nutzung eigener Datenquellen ist ein wichtiger Faktor, damit generative KI für das Unternehmen funktioniert. Aber das allein reicht nicht aus. Wie wir im weiteren Verlauf des Dokuments erörtern, müssen Entwickler auch überlegen, wie sie ihre Anwendung um ein informiertes großes Sprachmodell herum mit den richtigen Sicherheitskontrollen und in dem Umfang und der Leistung bereitstellen, die die Benutzer erwarten.

# Der Aufstieg der Vektoren und der Ähnlichkeitssuche

Um KI-Modelle mit unseren eigenen Daten zu versorgen, müssen wir diese zunächst in Vektoreinbettungen umwandeln. Diese Vektoren bieten mehrdimensionale numerische Kodierungen unserer Daten, die deren Muster, Beziehungen und Strukturen erfassen. Vektoreinbettungen verleihen unseren Daten eine semantische Bedeutung. Die Berechnung des Abstands zwischen Vektoren erleichtert es unseren Apps, die Beziehungen und Ähnlichkeiten zwischen verschiedenen Datenobjekten zu verstehen. Dies eröffnet unseren Daten eine ganze Reihe neuer Anwendungsmöglichkeiten, die wir weiter unten besprechen.

Daten in jedem digitalen Format und mit beliebiger Struktur – d. h. Text, Video, Audio, Bilder, Code, Tabellen – können in einen Vektor umgewandelt werden, indem sie mit einem geeigneten Vektoreinbettungsmodell verarbeitet werden. Zum Beispiel ist `text-embedding-ada-002` von OpenAI eines der beliebtesten Modelle für die Vektorisierung von Textinhalten. Das Schöne an Vektoreinbettungen ist, dass Daten, die unstrukturiert und daher für einen Computer völlig undurchsichtig sind, nun Bedeutung erhalten und ihre Strukturen über diese Einbettungen abgeleitet und dargestellt werden können. Das bedeutet, dass wir mit der Suche und Berechnung unstrukturierter Daten auf die gleiche Weise beginnen können, wie wir es schon immer mit strukturierten Geschäftsdaten konnten. Wenn man bedenkt, dass über 80 % der

Daten, die wir täglich erstellen, unstrukturiert sind, dann kann man einschätzen, wie transformativ die Vektorsuche in Kombination mit generativer KI wirklich ist.

Wie in Abbildung 1 unten dargestellt, werden unsere Daten, sobald sie in Vektoreinbettungen umgewandelt wurden, persistent gespeichert und in einem Vektorspeicher wie MongoDB Atlas Vector Search indexiert. Um ähnliche Vektoren abzurufen, wird der Speicher mit einem ANN-Algorithmus (Approximate Nearest Neighbor) abgefragt, um eine KNN-Suche (K Nearest Neighbor) mit einem Algorithmus wie „Hierarchical Navigable Small Worlds“ (HNSW) durchzuführen.

Die Abfrage dieser Vektoren ermöglicht es uns, Dinge mit Daten zu tun, die wir bisher nur mit teuren Data Science Kenntnissen und einer entsprechenden Infrastruktur erreichen konnten. Erstens können wir die Informationssuche und -entdeckung über den Abgleich von Schlüsselwörtern hinaus auf eine kontextbezogene semantische Suche ausdehnen, die in der Lage ist, aus dem Suchbegriff eines Benutzers die Bedeutung und Absicht abzuleiten. Zweitens können wir unsere eigenen Daten – kodiert als Vektoren – abrufen, um dem generativen KI-Modell den Kontext zu liefern, der notwendig ist, um zuverlässigere und genauere Ergebnisse zu erzeugen. Diese Outputs können umfassen:

- Natürliche Sprachverarbeitung (Natural Language Processing, NLP) für Aufgaben wie Chatbots und Fragenbeantwortung bis hin zur Textzusammenfassung und Stimmungsanalyse.
- Computer Vision und Audioverarbeitung zur Bildklassifizierung und Objekterkennung bis hin zur Spracherkennung und Übersetzung.
- Erstellung von Inhalten, einschließlich der Erstellung von textbasierter Dokumentation und SEO-optimierten Webseiten, Computercode oder der Umwandlung von Text in ein Bild oder ein Video.

# Vektorsuche und der LLM-Workflow

Abbildung 1 fasst den Workflow zusammen, der „Retrieval Augmented Generation“ für ein LLM ermöglicht.

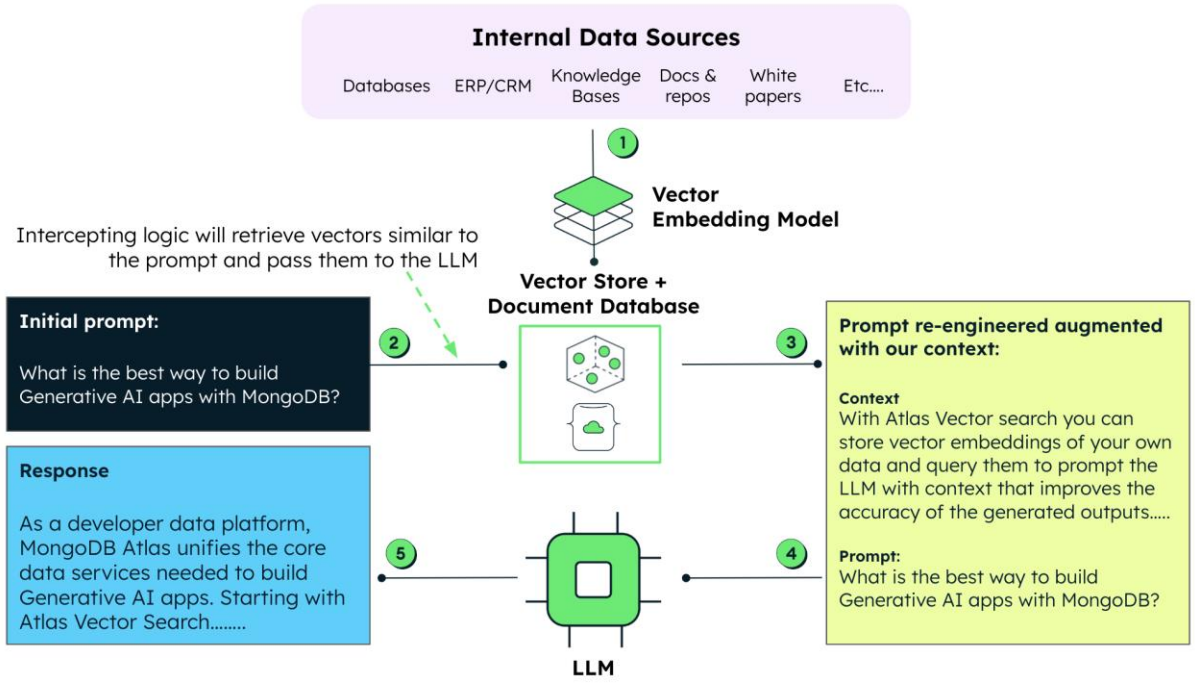


**Abbildung 1:** *Dynamische Kombination Ihrer benutzerdefinierten Daten mit dem LLM zur Generierung zuverlässiger und relevanter Ergebnisse*

Im Vorfeld werden unsere Daten durch ein Vektoreinbettungsmodell transformiert und in einem Vektorspeicher gespeichert. Idealerweise werden die Metadaten der Vektoren und die „gechunkten“ Rohdaten zusammen mit den Vektoren selbst in einer flexiblen Dokumentendatenbank gespeichert, die auch unsere regulären Anwendungsdaten speichert. Dadurch kann unsere Anwendung Daten auf verschiedene Arten abfragen, die Relevanz verbessern (z. B. neuere Daten höher bewerten) und stellt das Langzeitgedächtnis für das LLM zur Verfügung. Aufforderungen an das LLM werden von einer Logik abgefangen, die ähnliche Vektoren aus dem Vektorspeicher abruft. Diese werden dann verwendet, um die ursprüngliche Eingabeaufforderung zu überarbeiten. Die neue Eingabeaufforderung wird an das LLM gesendet, der den bereitgestellten Kontext nutzen kann, um anhand frischerer Daten qualitativ hochwertigere und genauere Antworten zu generieren.

Im weiteren Verlauf dieses Dokuments finden Sie Beispiele, die den obigen Arbeitsablauf veranschaulichen und zeigen, wie die daraus resultierenden Möglichkeiten auf verschiedene Klassen von Anwendungen angewendet werden können.

# Das Versprechen und die Realität einer lebendigen KI-Umgebung

Vektorspeicher sind Teil eines sich schnell entwickelnden Ökosystems von KI-unterstützenden Technologien, die von der Erstellung von Einbettungen bis hin zu

Prompt-Engineering, LLMs, Modellfeinabstimmung, Orchestrierung, Protokollierung, Infrastrukturautomatisierung und mehr reichen.

Innerhalb dieses Ökosystems gibt es eine Vielzahl interessanter und vielversprechender Projekte und Anbieter, mit denen Sie zusammenarbeiten können. Einige zeigen die „Kunst des Möglichen“ anhand von Demos und Prototypen. Aber die Angst, die Entscheidungsträger und Entwickler in Unternehmen haben, ist, wie leicht diese Prototypen für ihre spezifischen Geschäftsanforderungen angepasst werden können. Und ob einige der neueren Technologien wirklich in der Lage sind, die Produktionslast mit Zuverlässigkeit, Skalierbarkeit und Sicherheit zu bewältigen, Tag für Tag und in jeder Betriebsumgebung. Eine weitere Überlegung ist, wie die unternehmenseigenen Datenbanken integriert werden können, um echte, unverfälschte Geschäftsdaten in das Modell einzuspeisen.

Die KI-Umgebung existiert nicht in Isolation. All diese Technologien müssen in reale Anwendungen eingebettet werden, um für das Unternehmen wirklich nützlich zu sein. Vektorspeicher sind beispielsweise unerlässlich, um kontextsensitive generative KI und semantische Suche zu ermöglichen. Aber diese sind nur ein Teil einer umfassenderen Anwendung, die auch normale, nicht vektorisierte Geschäftsdaten verwalten muss.

Bei diesen Daten kann es sich um alles Mögliche handeln – Kundendatensätze, Bestellungen und Bestände, Geschäfte und Transaktionen, Angebote, geografische Koordinaten, Produktdetails und Preise, Zeitreihenmessungen und Sensormesswerte, Clickstreams und Social-Media-Feeds, Textbeschreibungen und vieles mehr.

Alle diese Daten müssen abgefragt werden, um die Anwendungsfunktionalität zu gewährleisten. Nicht nur, um ANNs zwischen Vektoren abzurufen, sondern auch, um regelmäßige Operationen auszuführen, wie z. B. das Abrufen bestimmter Datensätze, das Verarbeiten einer Flut von Datenaktualisierungen und das Ausführen anspruchsvoller Aggregationen und Transformationen zur Unterstützung der Analyseverarbeitung. Diese Abfragen unterstützen Anwendungsfunktionen außerhalb von generativen KI-Anwendungsfällen. Aber sie werden noch wichtiger, wenn wir sie zusammen mit kontextabhängigen Aufforderungen an unsere Modelle verwenden können, um die Genauigkeit und Relevanz der Ergebnisse des generativen KI-Modells zu verbessern.

Neben der Arbeit mit unseren Anwendungsdaten und Vektoreinbettungen müssen wir auch die nicht-funktionalen Dinge erledigen – die Einhaltung von SLAs für Betriebszeit, Leistung und Skalierbarkeit, die Integration neuer Funktionen, die Sicherung und das Backup von Daten sowie deren Überprüfung. Manches davon kann langweilig klingen. Das heißt, bis es nicht mehr funktioniert. Dann ist es plötzlich nicht mehr langweilig …

Wenn Sie die Technologien für neue KI-gestützte Erlebnisse zusammenführen und in Ihre Anwendungen integrieren, besteht die Gefahr, dass ein Sammelsurium von

Einzelprodukten und Komplexität entsteht, das Ihren Teams einen enormen Mehraufwand beschert. All diese Herausforderungen summieren sich zu fragmentierten und ineffizienten Entwicklererfahrungen, einer Vielzahl von Betriebs- und Sicherheitsmodellen, die es zu bewältigen gilt, einer Unmenge von Datenkämpfen und Integrationsarbeiten und einer Menge von Datenduplikaten. All dies verlangsamt die Geschwindigkeit, mit der Sie Ihre neuen KI-gestützten Erfahrungen auf den Markt bringen und erhöht gleichzeitig Ihre Kosten und Risiken.

Mit einer Entwicklerdatenplattform, die auf MongoDB Atlas aufbaut, gibt es einen besseren Weg.

# Eine Datenplattform für Entwickler: der intelligente Weg zur Entwicklung intelligenter Anwendungen

Die MongoDB Datenplattform für Entwickler, die auf MongoDB Atlas aufbaut, vereint operative, analytische und generative KI-Datendienste, um die Entwicklung intelligenter Anwendungen zu optimieren. Wie auch immer Sie KI nutzen – vom Trainieren und Bereitstellen Ihrer eigenen maschinellen Lernmodelle bis hin zum Einbetten der neuesten generativen KI in Ihre Anwendungen – Atlas ist ein wichtiger Bestandteil Ihres Stacks. Vom Prototyp bis zur Produktion können Sie mit Atlas sicherstellen, dass Ihre Anwendungen auf den aktuellsten Betriebsdaten basieren und gleichzeitig die Skalierung, Sicherheit und Leistung bieten, die Benutzer erwarten.

Das Herzstück von MongoDB Atlas ist das flexible Dokumentdatenmodell und die entwicklernahe Abfrage-API. Zusammen ermöglichen sie es Ihren Entwicklern, das Innovationstempo drastisch zu beschleunigen, die Konkurrenz zu überholen und die neuen Marktchancen zu nutzen, die sich durch generative KI ergeben.

Dokumente sind der beste Weg für Entwickler, mit Daten zu arbeiten, da sie auf Objekte im Code abgebildet werden und somit intuitiv und einfach zu verstehen sind. Dokumente können Daten jeglicher Struktur modellieren – von der großen Vielfalt regulärer Anwendungsdaten, die wir bereits besprochen haben, bis hin zu Vektoreinbettungen, die aus mehreren tausend Dimensionen bestehen. Jede dieser Strukturen kann jederzeit geändert werden, um das Hinzufügen neuer Datentypen und Anwendungsfunktionen zu unterstützen. Dokumente geben Ihnen die Flexibilität, diese Daten auf eine Art und Weise zu rationalisieren und zu nutzen, wie es bei herkömmlichen tabellarischen Datenmodellen relationaler Datenbanken nicht möglich ist.

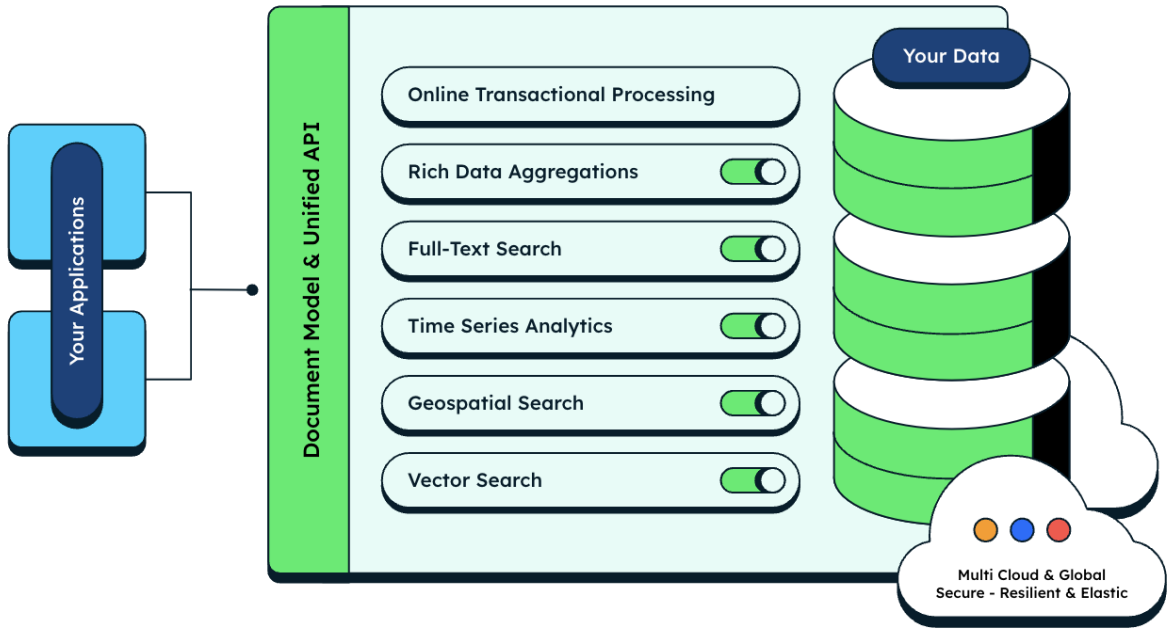


**Abbildung 2:** *MongoDB Atlas integriert die Datendienste, die Sie benötigen, um KI in Ihre Anwendungen zu bringen*

In Verbindung mit dem Dokumentenmodell bietet die MongoDB Query API Entwicklern eine einheitliche und konsistente Möglichkeit, mit Daten in jedem Datendienst zu arbeiten. Von einfachen CRUD-Operationen über die Suche nach Schlüsselwörtern und Vektorähnlichkeiten bis hin zu ausgefeilten Aggregationspipelines für Analysen und Stream Processing bietet die MongoDB Abfrage-API Entwicklern die Flexibilität, Daten je nach Bedarf der Anwendung abzufragen und zu berechnen. Im Kontext von generativer KI bietet dies äußerst flexible und leistungsstarke Möglichkeiten, zusätzliche Filter für vektorbasierte Abfragen zu definieren, zum Beispiel:

- Kombiniert mit Metadaten zum Filtern: „Finde Inhalte, die der Suchanfrage des Benutzers entsprechen, aber nur Inhalte, die in den Jahren X, Y und Z veröffentlicht wurden."
- Kombiniert mit Aggregationen: „Finde alle Bilder, die dem abgefragten Bild ähnlich sind, und gruppiere sie nach der Fotografen-ID."
- Kombiniert mit der geografischen Suche: „Finden für mich Immobilienangebote für Häuser, die dem Haus auf diesem Foto ähnlich sind und sich in einem Umkreis von N Meilen von meinem Standort befinden."

Keine andere Datenbank ist in der Lage, eine so große Bandbreite an Abfragefunktionen in einer zentralen, vereinheitlichten Abfrageumgebung anzubieten. Dies ermöglicht es Entwicklern, Endbenutzerfunktionen einfacher und mit weniger Komplexität zu entwickeln. Entwickler müssen Abfrageergebnisse aus mehreren Datenbanken nicht mehr manuell zusammenfügen – was einen komplexen, fehleranfälligen, kostspieligen und langsamen Prozess erfordert. Gleichzeitig bleibt Ihr technologischer Fußabdruck kompakt und agil.

*„MongoDB speicherte bereits Metadaten über Artefakte in unserem System. Mit der Einführung von Atlas Vector Search verfügen wir nun über eine umfassende Datenbank für Vektor-Metadaten, die sich seit einem Jahrzehnt bewährt hat und unsere dichten Suchanforderungen erfüllt. Wir müssen keine neue Datenbank einrichten, die wir verwalten und erlernen müssten. Unsere Vektoren und Artefakt-Metadaten können direkt nebeneinander gespeichert werden.“*

Pierce Lamb, Senior Software Engineer im Team für Daten und maschinelles Lernen bei VISO TRUST.

# Show, don't tell. Generative KI-verbesserte Apps auf einer Entwicklerdatenplattform

Wir werden uns auf drei beliebte Anwendungsfälle konzentrieren, um zu zeigen, wie Entwickler MongoDB Atlas nutzen, um KI-angereicherte Apps zu erstellen:

- Chatbot und Question-Answering (Q&A) für Kunden-Self-Service.
- Erweiterte E-Commerce-Suche und Benutzerempfehlungen.
- Analyse und Generierung von Rich Media (multimodal).

Jedes dieser Beispiele basiert auf generativer KI und fortschrittlicher semantischer Suche, um beeindruckende Benutzererfahrungen zu schaffen und Fähigkeiten freizusetzen, die für die meisten Unternehmen bisher unerreichbar waren. Um jedoch wirklich transformativ zu sein, müssen diese KI-Verbesserungen als Teil einer größeren Anwendung bereitgestellt werden, die selbst wichtige Geschäftsfunktionen bereitstellt.

Wir gehen jeden Anwendungsfall der Reihe nach durch und zeigen ein architektonisches Entwurfsmuster, das ihn unterstützt, sowie die entsprechenden Funktionen von MongoDB Atlas.

## Chatbot und Q&A für den Kunden-Selfservice

MongoDB ist das Herzstück vieler Kundensupport-Anwendungen. Denn das flexible Datenmodell von MongoDB macht es einfach, eine zentrale 360-Grad-Sicht auf den Kunden zu erstellen. Dies geschieht durch die dynamische Aufnahme unterschiedlicher und sich schnell ändernder Kundendaten aus den unzähligen, in den meisten Unternehmen typischen, isolierten Backend-Quellsystemen. Die einheitliche, konsolidierte Echtzeit-Kundenansicht, die von MongoDB unterstützt wird, ist daher die ideale Plattform, auf der wir Chatbot- und Q&A-Hilfsfunktionen für den Kunden-Selfservice trainieren und bereitstellen können.

In unserem Beispiel in Abbildung 2 wird die in MongoDB gespeicherte Kundendatenbank als JSON-Datei in ein Einbettungsmodell exportiert, das die Daten (mithilfe von Tools wie LangChain oder LlamaIndex) fragmentiert und daraus Vektoreinbettungen erstellt. Andere interne Datenquellen wie Wissensdatenbanken und Dokumentationen können ebenfalls für die Verwendung in der App vektorisiert werden. Die Daten werden dann wieder in die MongoDB Datenbank importiert.

Wir müssen sicherstellen, dass unsere Vektoren ständig mit den neuesten Kundendaten aktualisiert werden, daher verwenden wir Atlas Triggers, um alle Datenänderungen in unserer Einzelansicht zu überwachen. Sobald neue Kundendatensätze eingefügt oder bestehende Datensätze in der Datenbank aktualisiert werden, ruft Atlas Triggers die API des Einbettungsmodells auf, um die entsprechenden Vektoren zu generieren und sie zurück in Atlas zu laden.

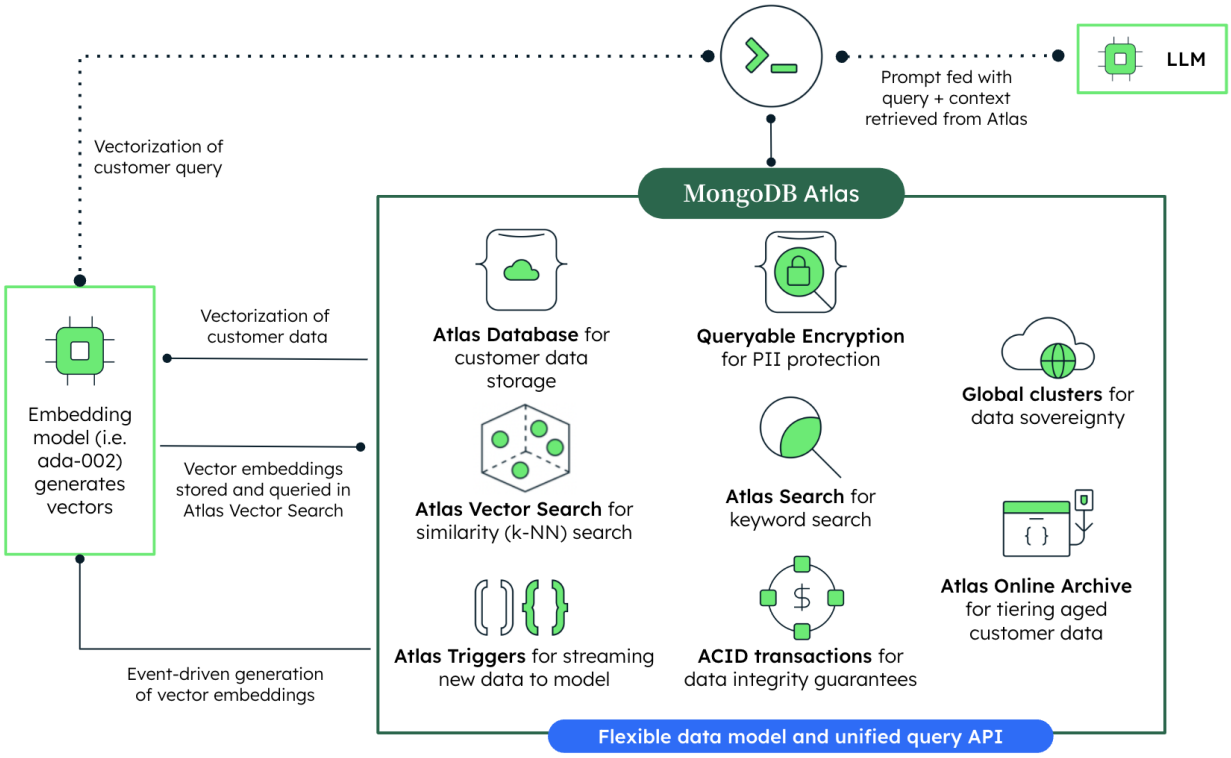


**Abbildung 3:** *Chatbot und Q&A generative KI-Funktionen in einer Kunden Self-Service-Anwendung auf Basis von MongoDB Atlas*

Mit Atlas nutzen Entwickler das flexible Datenmodell von MongoDB. Sie können die Quellkundendaten, Metadaten und Chunks zusammen mit den Vektoreinbettungen speichern, alles synchronisiert und Seite an Seite in einer einzigen Speicherebene, auf die über eine einzige Abfrage-API und einen einzigen Treiber zugegriffen werden kann.

Abfragen können Daten effizient filtern, indem sie die indizierten Vektoren neben den Schlüsselwortindizes der regulären Felder in Ihren Dokumenten verwenden. Diese Integration bedeutet, dass die App ein viel breiteres Spektrum an Benutzerfunktionen mit geringerem Aufwand für die Entwickler unterstützen kann:

- Atlas Vector Search liefert übereinstimmende Dokumente, indem es eine Ähnlichkeitssuche in den indizierten Einbettungsdaten durchführt. Um das Risiko zu verringern, dass veraltete Daten zurückgegeben werden, können unsere Abfragen die Metadaten eines Vektors – wie das in der Atlas-Datenbank gespeicherte „Erstellungsdatum“ –verwenden, um ältere Inhalte herauszufiltern.
- Atlas Search liefert Ergebnisse, die auf übereinstimmenden Schlüsselwörtern in den Quell- und „Chunked“-Kundendaten basieren. Sie verwendet Funktionen wie die unscharfe Suche, um Tippfehler in Benutzereingaben zu korrigieren und die automatische Vervollständigung, um vorgeschlagene Suchbegriffe zu liefern. Außerdem verwendet sie Indexüberschneidungen, um komplexe Ad-hoc-Abfragen der Kundendaten effizient zu bearbeiten.

Queries an die Atlas Datenbank, Vector Search, und Atlas Search verwenden alle dieselbe Abfrageoberfläche und denselben Treiber, was den Arbeitsablauf der Entwickler erheblich vereinfacht. Die aus MongoDB Atlas abgerufenen Daten werden als Kontext zur Erweiterung der Eingabeaufforderung an das LLM geliefert, sodass es relevante Antworten auf Chats und Fragen generieren kann. Der Kontext und die Aufforderungen sowie alle damit verbundenen Argumentationsschritte, die zur Beantwortung komplexer Fragen verwendet werden, werden in Atlas gespeichert, sodass der LLM über ein Langzeitgedächtnis verfügt und seine Ergebnisse kontinuierlich verbessert.

Kundendaten gehören zu den wertvollsten Daten, die ein Unternehmen verwaltet. Während generative KI uns dabei hilft, die Art und Weise, wie wir unsere Kunden betreuen, zu verbessern, bleibt der Schutz ihrer Daten oberstes Gebot. Atlas bietet eine Reihe von Funktionen, die uns dabei helfen, sodass sich die Entwickler auf KI-gestützte Funktionen konzentrieren können:

- Konvergente Infrastruktur für Datenspeicherung, Abfragen und Analysen, Stichwortsuche und Vektorsuche. Durch diese Vereinheitlichung hinter einer einzigen API und einem einzigen Datenmodell wird die Anzahl der beweglichen Teile, die Entwickler integrieren und mit denen sie arbeiten müssen, drastisch reduziert.
- Queryable Encryption ist eine branchenweit führende Lösung zur Sicherung von Kundendaten. MongoDB Treiber verschlüsseln sensible Datenfelder clientseitig, wobei die Datenbank immer nur mit ihnen als vollständig randomisierte verschlüsselte Daten arbeitet. Selbst wenn die Daten verschlüsselt sind, können Anwendungen aussagekräftige Abfragen durchführen, ohne die Daten in der Datenbank entschlüsseln zu müssen. Beachten Sie, dass in der Regel nur die Felder, die die sensibelsten Daten zur eindeutigen Identifizierung einer Person enthalten, wie z. B. die Sozialversicherungsnummer, mit Queryable Encryption geschützt werden. Die Suche kann daher in den verbleibenden Klartextfeldern durchgeführt werden.

- Multi-Dokument-ACID-Transaktionen in der Atlas-Datenbank garantieren die Integrität unserer Kundendaten, wenn sie von der Anwendung aufgerufen und geändert werden.
- Mit Atlas Global Clusters können Kundendaten an die Region gebunden werden, in der sie sich befinden, was die modernen Vorschriften zur Datenhoheit erfüllt.
- Atlas Online Archive bietet eine vollständige Verwaltung des Lebenszyklus von Daten. Der Service lagert gealterte Kundendaten automatisch aus aktiven Datenbanken in einen kostengünstigeren Cloud-Objektspeicher aus, wobei die Daten für Abfragen zugänglich bleiben. Dies ist wichtig für Kundendaten, die in Anwendungen verwaltet werden, die in regulierten Branchen tätig sind, wo sie mehrere Jahre lang aufbewahrt werden und zugänglich sein müssen.
- Kundendaten werden durch Backups und Point-in-Time-Wiederherstellung vor Beschädigung und Ransomware geschützt.

Atlas wird auf den wichtigsten Hyperscaler-Clouds vollständig für Sie verwaltet und durch ein SLA mit einer Verfügbarkeit von 99,995 % abgesichert.

## Erweiterte E-Commerce-Suche und Empfehlungen

E-Commerce-Produktkataloge sind ein häufiger Anwendungsfall für MongoDB:

- Die Vielfalt der verschiedenen Produkte und ihrer Attribute spiegelt sich natürlich im flexiblen Dokumentdatenmodell von MongoDB wider.
- Die verteilte Architektur von Atlas mit elastischer Skalierung ermöglicht es Entwicklern, die Datenbankkapazität dynamisch an den Bedarf der Anwendung anzupassen (z. B. für saisonale Einkäufe und Verkaufsangebote).
- Mit Atlas Search ermöglichen Keyword-Matching-Funktionen wie Fuzzy-Suche, Autovervollständigung, Facettierung, Hervorhebung und benutzerdefiniertes Scoring den Käufern ein schnelles Durchsuchen und Navigieren im Produktkatalog, wodurch die Klickraten (CTRs) und die Kaufkonversion gesteigert werden.

Die Suche nach Schlüsselwörtern beruht jedoch auf der Übereinstimmung bestimmter Wörter in indizierten Textfeldern, um relevante Ergebnisse zu liefern. Ohne eine umfangreiche und mühsame Zuordnung von Synonymen (z. B. von Fahrrädern zu Rädern oder von Turnschuhen zu Sportschuhen) werden Benutzer schnell frustriert sein, wenn ihre Suchanfragen keine relevanten Produkte ergeben. Diese Frustration führt zu Umsatzeinbußen und einem geschädigten Ruf der Marke.

Eine weitere Herausforderung ist die Bereitstellung von Empfehlungen für Benutzer. Entwickler müssen entweder komplexe regelbasierte Engines entwickeln oder auf spezialisierte und knappe Data Science Ressourcen zurückgreifen. In der Regel müssen die Daten zunächst per ETL (Extrahieren, Transformieren, Laden) aus der operativen Datenbank in ein Offline Data Warehouse oder einen Data Lake übertragen werden. Erst dann können traditionelle analytische KI-Modelle eine Reihe von Empfehlungen

generieren, die dann wieder in die operative Datenbank geladen werden müssen. Das Verfahren ist komplex, teuer und erzeugt Empfehlungen, die unmittelbar veraltet sind, da sie nicht das letzte Surfverhalten oder die letzten Einkäufe des Benutzers widerspiegeln.

Die Erweiterung unseres Produktkatalogs mit Vektoreinbettungen beseitigt diese Herausforderungen. Vektoren verleihen den Produkten in unserem Katalog eine semantische Bedeutung, sodass Ähnlichkeiten und die Beziehungen zwischen den Produkten leicht zu erkennen sind. Auf diese Weise können Händler den Nutzern relevante und verwandte Produkte mit deutlich weniger Aufwand, Komplexität und Kosten präsentieren. Gängige Suchbegriffe können in MongoDB Atlas zwischengespeichert werden, sodass Benutzern relevante Ergebnisse schneller zur Verfügung stehen.

Die Ausweitung der Vektorisierung auf Kundendaten – wie bereits in der Kunden-Self-Service-App demonstriert – ermöglicht es uns, noch ausgefeiltere Empfehlungen zu erstellen, indem wir die Suche nach Produkt- und Kundenähnlichkeiten kombinieren, um die Vorschläge zu verfeinern.

Abbildung 4 zeigt ein übergeordnetes Designmuster für die erweiterte Suche und Empfehlungen. Die Erstellung und Pflege unserer Vektoreinbettungen folgt dem gleichen Arbeitsablauf, der bereits für Chatbots und Q&A in unserer Kunden Self-Service-Anwendung beschrieben wurde.

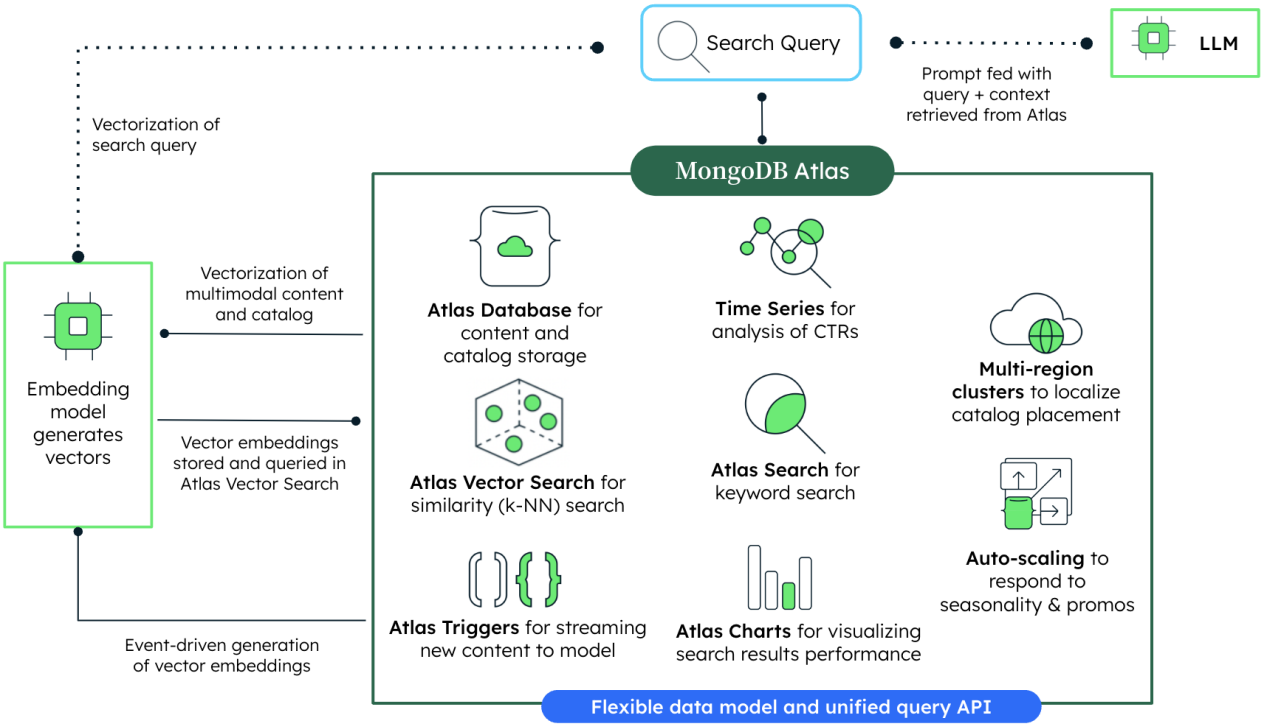


**Abbildung 4:** *Die erweiterte semantische Suche in unserem Produktkatalog führt zu höheren Umsätzen und Upsell*

Es ist leicht zu erkennen, wie die Vektorsuche die Produktsuche und -empfehlungen dramatisch verbessert. Die Integration eines LLM bringt diese Erfahrung noch weiter. Jetzt können Kunden live Fragen stellen und erhalten sofortige Antworten zu den Produkten, die sie bewerten, was den Kaufzyklus beschleunigt.

Vertriebsmitarbeiter können das LLM auch für eine Reihe von Aufgaben nutzen, die zuvor mühsam waren, und so noch kreativere Wege zur Kundenbindung entwickeln. Mit dem LLM können beispielsweise verschiedene Varianten von Produkttexten und SEO-Schlüsselwörtern erstellt werden, die dann in A/B-Tests getestet werden können, um festzustellen, welche davon zu mehr Konversionen führen. Das LLM könnte dazu verwendet werden, mehrere Nutzerbewertungen zusammenzufassen und daraus Stimmungen abzuleiten, um das Feedback zu synthetisieren, das in die Produktplanung einfließt.

Unternehmen können Atlas nutzen, um den gesamten E-Commerce-Lebenszyklus zu verwalten. Neben der Nutzung von KI, um unsere Sucherfahrung intelligenter und vorausschauender zu gestalten, können Unternehmen die Klickraten und Verkaufszahlen von Suchergebnissen verfolgen. Time-Series Collections können hochfrequente und umfangreiche Klickströme von User Sessions effizient aufnehmen und speichern. Diese Daten stehen für Analysen zur Messung der Suchleistung zur Verfügung, einschließlich Live-Visualisierungen der Ergebnisse mit Atlas Charts.
Mit diesen Erkenntnissen können Vertriebsmitarbeiter die Produktdaten und die Relevanzbewertung kontinuierlich abstimmen und optimieren, um die Umsätze auf der E-Commerce-Website zu maximieren.

## Rich Media (multimodale) Analyse und Generierung

Die normale Textsuche ist mit der herkömmlichen Stichwortsuche gut bedient.
Die Arbeit mit reichhaltigeren Medien (manchmal auch multimodal genannt) wie Bildern, Sprache und Video erfordert jedoch hochkomplexe Data Science Technologien und Fähigkeiten. Bis jetzt.

Wie bereits erwähnt, kann jeder digitale Inhalt mit einem geeigneten Vektoreinbettungsmodell vektorisiert werden. KI-Hubs wie Hugging Face und die von den Cloud-Hyperscalern bieten eine Fülle von Modellen, die auf verschiedene Inhaltsmodalitäten abgestimmt sind. Die Einbettungen aus diesen Modellen können in Atlas Vector Search gespeichert werden, um eine ganze Reihe neuer Funktionen zu ermöglichen. Wie bereits erwähnt, sind die Generierung von Bildern aus Text, die Transkription von Videos für die Spracherkennung und Stimmungsanalyse, die Klassifizierung von Bildern und die Erkennung von Objekten nur einige Beispiele für die Möglichkeiten. Vektoren aus verschiedenen Medien können kombiniert werden – zum Beispiel durch den Vergleich einer Text- und Bildeinbettung, um zu prüfen, ob ein bestimmter Satz ein Bild genau beschreibt.

Diese multimodale Funktionalität kann für eine Reihe von Anwendungsfällen genutzt werden. Zum Beispiel die Anreicherung von Produktkatalogen, wie oben beschrieben, oder die Verbesserung der Entdeckung von Bildern und Videos durch Analyse. Dadurch können Design-, Herstellungs- und Veröffentlichungsprozesse optimiert oder völlig neue Klassen von Anwendungen in Bereichen wie Sicherheit und Überwachung oder Augmented Reality (AR) geschaffen werden.

Das architektonische Designmuster und die Fähigkeiten von MongoDB Atlas, die oben für die erweiterte E-Commerce-Suche und Empfehlungen beschrieben wurden, gelten auch für die Erstellung multimodaler Inhalte.

# MongoDB Vector Search in Aktion

MongoDB wird bereits in großem Umfang für traditionelle KI-Anwendungsfälle eingesetzt. Continental hat MongoDB für die Feature-Engineering-Plattform seiner Vision Zero-Initiative für autonomes Fahren ausgewählt. Sowohl Bosch als auch Telefonica verwenden MongoDB in ihren KI-gestützten IoT-Plattformen. Kronos handelt täglich mit Kryptowährungen im Wert von Milliarden von Dollar und verwendet dazu ML-Modelle, die mit Daten aus MongoDB konfiguriert und erstellt wurden. Iguazio verwendet MongoDB als Persistenzschicht für seine Data Science- und MLOps-Plattform, während H2O.ai und Featureform MongoDB als Feature-Store in ihren jeweiligen Plattformen unterstützen.

Auf dieser Grundlage wird MongoDB Atlas bereits heute in einer Reihe von Anwendungen eingesetzt, die die Grenzen dessen, was mit generativer KI möglich ist, verschieben. Werfen Sie einen Blick auf unsere Seite mit Fallstudien, um mehr über die Bandbreite der Anwendungsfälle zu erfahren, für die MongoDB Atlas eingesetzt wird. Eine Auswahl an konkreten Beispielen:

- Ada: unterstützt Unternehmen wie Meta, ATT und Verizon dabei, ihre Kunden durch KI-gestützte Automatisierung und dialogorientierte KI besser zu betreuen.
- ExTrac: identifiziert und klassifiziert aufkommende physische und digitale Risiken anhand der Analyse von Echtzeit-Datenströmen.
- Eni: erschließt geologische Daten und macht sie für eine bessere Entscheidungsfindung nutzbar und beschleunigt den Weg des Unternehmens hin zu Net Zero.
- Inovaare: überwacht, extrahiert und klassifiziert kontinuierlich Daten über den gesamten Lebenszyklus des Gesundheitswesens für die Berichterstellung zur Einhaltung gesetzlicher Vorschriften, für Audits und Risikobewertungen.
- Source Digital: erzielt eine 7-fache Kostenreduzierung nach der Migration von PostgreSQL zu MongDB Atlas für seine Videoerkennungsplattform.

- Catylex: extrahiert, klassifiziert und analysiert automatisch Vertragsbedingungen, um Rechte, Pflichten und Risiken zu identifizieren
- Robust Intelligence: schützt Large Language Models (LLMs) in der Produktion durch die Validierung von Eingaben und Ausgaben in Echtzeit mit seinem KI-Firewall-Angebot.
- Potion: regeneriert Video- und Audiostreams mithilfe von benutzerdefinierten Bild- und Audiomodellen.


**Abbildung 5:** *Retool State of AI Survey – die branchenweit besten Vektordatenbanken*

Die Beliebtheit von MongoDB bei KI-Entwicklern spiegelt sich auch darin wider, dass der Software-Tool-Anbieter Retool in seinem State of AI Survey feststellte, dass MongoDB Atlas Vector Search:

1. den höchsten Net Promoter Score (NPS) aller untersuchten Vektordatenbanken aufweist.
2. innerhalb weniger Monate nach seiner Markteinführung zur am zweithäufigsten verwendeten Vektordatenbank aufgestiegen ist und damit vor alternativen Lösungen liegt, die bereits seit Jahren auf dem Markt sind.

*„Atlas Vector Search ist robust, kostengünstig und blitzschnell!“*

Saravana Kumar, CEO, Kovai, über die Entwicklung des KI-Assistenten seines Unternehmens.

# Erste Schritte

Ganz gleich, ob Sie in einem Startup oder in einem Unternehmen die Technologie von morgen entwickeln, mit MongoDB Atlas können Sie Folgendes erreichen:

- Beschleunigung der Entwicklung Ihrer mit generativer KI erweiterten Anwendungen, die auf tatsächlichen Betriebsdaten beruhen.
- Vereinfachung Ihres Technologie-Stacks, indem Sie eine zentrale Plattform nutzen, die es Ihrer App ermöglicht, operative Daten und Vektoreinbettungen am selben Ort zu speichern, mit serverlosen Funktionen auf Änderungen in den Quelldaten zu reagieren und über mehrere Datenmodalitäten hinweg zu suchen – und so die Relevanz und Genauigkeit der von der App generierten Antworten zu verbessern.
- Weiterentwicklung Ihrer generativen, mit KI erweiterten Apps mit der Flexibilität des Dokumentenmodells unter Beibehaltung einer einfachen, eleganten Entwicklererfahrung.
- Nahtlose Integration führender KI-Services und -Systeme wie die Hyperscaler und Open Source LLMs und Frameworks, um in dynamischen Märkten wettbewerbsfähig zu bleiben.
- Entwicklung von mit GenKI erweiterten Anwendungen auf einer leistungsstarken, hoch skalierbaren operativen Datenbank, die seit einem Jahrzehnt für eine Vielzahl von KI-Anwendungsfällen validiert wurde.

Weitere Informationen zum Erstellen von KI-gestützten Apps mit MongoDB finden Sie in unserem KI/ML-Ressourcenzentrum.

Der beste Weg für Entwickler ist, sich für ein Konto bei MongoDB Atlas anzumelden. Von dort aus können sie eine kostenlose MongoDB Instanz mit der Atlas Datenbank, Atlas Vector Search und Atlas Search erstellen, ihre eigenen Daten oder unsere Beispieldatensätze laden und erkunden, was innerhalb der Plattform möglich ist.

## Absicherungserklärung

Entwicklung, Freigabe und zeitliche Planung der beschriebenen Funktionen oder Features für unsere Produkte liegen in unserem alleinigen Ermessen. Diese Informationen dienen lediglich dazu, unsere allgemeine Produktrichtung zu umreißen, und sollten nicht als Grundlage für eine Kaufentscheidung herangezogen werden. Dies ist auch keine Selbstverpflichtung, Zusage oder rechtliche Verpflichtung zur Lieferung von Material, Code oder Funktionalität irgendwelcher Art.

US 866-237-8815 • INTL +1-650-440-4474 • info@mongodb.com.