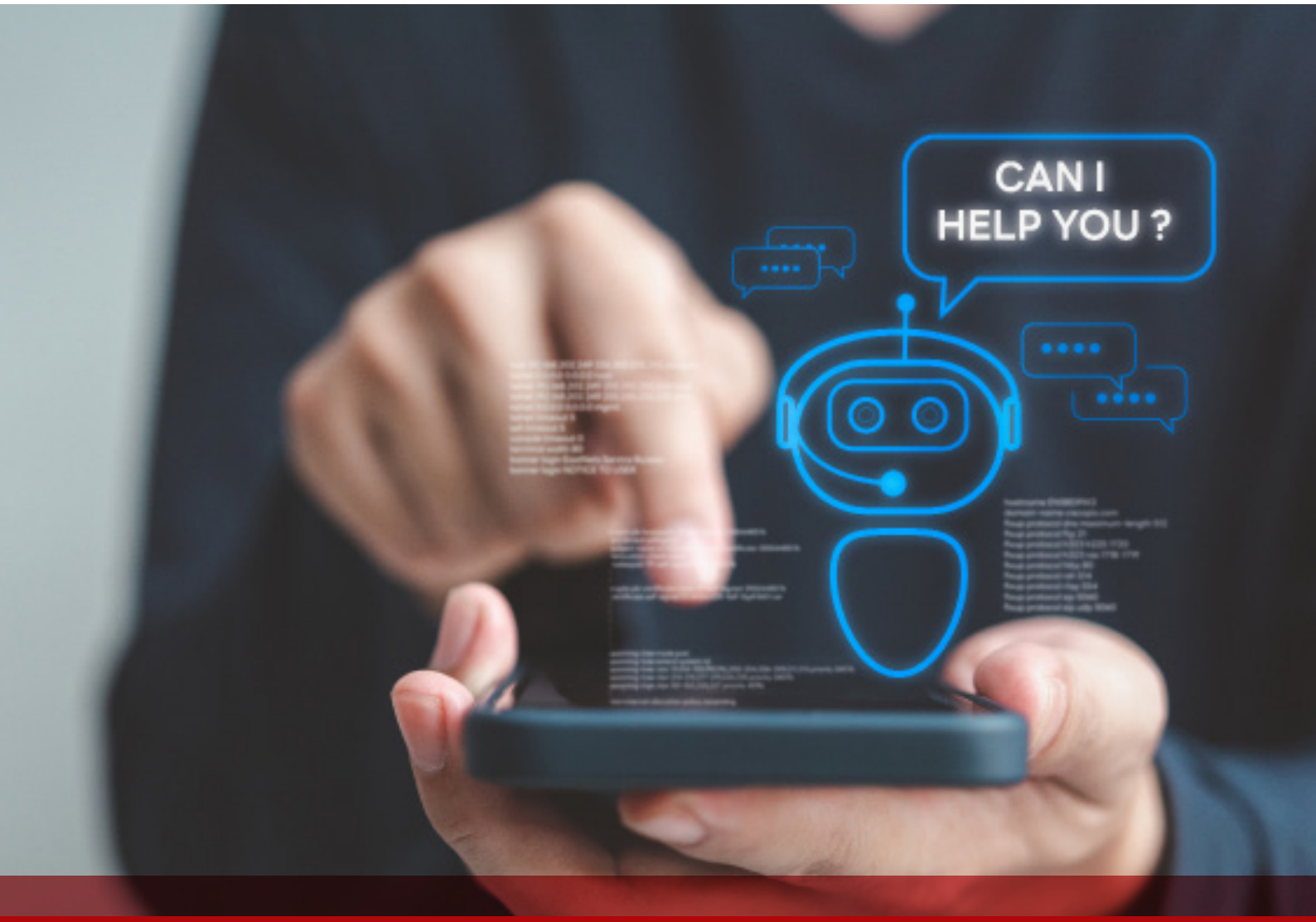
Mittelstand-Digital
Zentrum
Augsburg



Kompaktwissen KI

RETRIEVAL AUGMENTED GENERATION (RAG) IN DER PRAXIS

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

Mittelstand-
Digital

aufgrund eines Beschlusses
des Deutschen Bundestages

INHALT

IMPRESSUM

Verleger:
Fraunhofer-Institut für Gießerei-, Composite und Verarbeitungstechnik IGCV

Am Technologiezentrum 2 • 86159 Augsburg
Als rechtlich nicht selbstständiges Institut der Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e. V.

Hansastraße 27c • 80686 München
Tel.: 0821 90678-0 • Fax: 0821 90678-40
E-Mail: info@igcv.fraunhofer.de

Rechtsform:
Das Fraunhofer-Institut für Gießerei-, Composite und Verarbeitungstechnik IGCV ist ein rechtlich nicht selbstständiges Institut der Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e. V.

Vertretung:
Präsident des Vorstandes: Prof. Dr.-Ing. Holger Hanselka

Text und Redaktion:
Sebastian Maier, Fraunhofer IGCV
Florian Doll, Fraunhofer IGCV
Maria Bachmaier, VDMA Bayern

Bildquellen:
© khunkornStudio / Adobe Stock (Titelseite)
© Unicorns / Adobe Stock (Seite 7)
© DC - Studio / Adobe Stock (Seite 10)
© Dusan Petkovic / Adobe Stock (Seite 13)
© vectorfusionart / Adobe Stock (Seite 18)

Stand:
Juni 2026

1 EINLEITUNG

Künstliche Intelligenz, insbesondere Große Sprachmodelle (Large Language Models, kurz LLMs), haben die Art und Weise, wie wir mit Texten arbeiten, revolutioniert. Doch so beeindruckend diese Modelle sind, für den Mittelstand haben sie oft ein entscheidendes Problem: Sie wissen nichts über Ihr Unternehmen. Sie kennen weder Ihre internen Verträge noch Ihre spezifischen Produkthandbücher oder den aktuellen Lagerbestand. Zudem neigen sie zu „Halluzinationen". Sie erfinden Fakten, wenn sie die Antwort nicht wissen.

Hier setzt **Retrieval Augmented Generation (RAG)** an. RAG ist die Brücke zwischen der mächtigen Sprachfähigkeit einer KI und dem exklusiven Wissen Ihres Unternehmens. Anstatt ein Modell teuer neu zu trainieren, gibt man ihm die Fähigkeit, in Ihren eigenen Daten „nachzuschlagen", bevor es antwortet.

Dieser Leitfaden soll Geschäftsführenden und IT-Verantwortlichen im Mittelstand zeigen, wie RAG funktioniert, warum es die Schlüsseltechnologie für den produktiven KI-Einsatz ist und wie man ein solches System von der Rohdatei bis zum Chatbot technisch umsetzt.

2 BEGRIFFE UND FUNKTIONSWEISE IM KONTEXT VON RAG

Um das Potenzial von Retrieval Augmented Generation für das eigene Unternehmen zu bewerten, ist ein fundiertes Verständnis der zugrundeliegenden Mechanismen notwendig. Dieses Kapitel erläutert, was RAG technisch ausmacht, wie es sich von anderen KI-Methoden unterscheidet und welche Bausteine für die Funktion unerlässlich sind.

2.1 WAS IST RAG?

Retrieval Augmented Generation (dt.: Abfrage-erweiterte Generierung) ist eine Hybrid-Architektur, die die sprachliche Kompetenz eines Großen Sprachmodells (LLM) mit dem fachlichen Tiefenwissen einer externen Datenbank kombiniert.

Ein herkömmliches KI-Modell ist ein statisches System. Sein Wissen ist auf den Zeitpunkt seines Trainings eingefroren („Knowledge Cutoff"). Fragt man es nach aktuellen Umsatzzahlen oder einer internen Richtlinie von letzter Woche, kann es diese Information nicht liefern oder es „halluziniert", d. h. es erfindet eine plausibel klingende, aber faktisch falsche Antwort.

RAG löst dieses Problem, indem es dem Modell ein dynamisches Gedächtnis zur Seite stellt. Bevor die KI eine Antwort formuliert, führt sie eine Recherche in den bereitgestellten Unternehmensdaten durch. Das Modell fungiert nicht mehr als allwissendes Orakel, sondern als intelligenter Lese-Assistent, der Informationen aus den Dokumenten extrahiert, zusammenfasst und verständlich aufbereitet.

2.2 RAG VS. FINE-TUNING

Ein häufiges Missverständnis im Mittelstand ist die Annahme, man müsse eine KI „trainieren", damit sie firmenspezifisches Wissen lernt. Hier muss zwischen Fine-Tuning und RAG unterschieden werden:

Fine-Tuning (Training)

Hierbei wird das Gehirn der KI dauerhaft verändert. Das Modell lernt neue Muster, z. B. einen spezifischen Jargon oder eine besondere Art zu programmieren. Es ist vergleichbar mit einem Studierenden, der eine neue Sprache lernt. Das Hinzufügen von neuem Faktenwissen ist hier jedoch mühsam, teuer und erfordert bei jeder Änderung ein erneutes Training.

RAG (Kontext)

Das Modell bleibt unverändert. Man gibt dem Studierenden lediglich ein aktuelles Fachbuch an die Hand, in dem er die Lösung nachschlagen kann. Ändern sich die Informationen (z. B. neue Preise), tauscht man einfach die Seite im Buch aus.

Das bedeutet zusammengefasst, Fine-Tuning wird vor allem genutzt, wenn die KI lernen soll, wie sie sprechen soll (Stil, Formate). RAG wird hingegen angewendet, wenn die KI wissen muss, welcher Sachstand aktuell gilt (Fakten, Daten).

2.3 DIE DREI PHASEN DES PROZESSES

Der RAG-Workflow lässt sich in drei logische Schritte unterteilen, die in Bruchteilen einer Sekunde ablaufen:

Retrieval

Der Nutzende stellt eine Frage (z. B. „Wie hoch ist die Toleranz bei Bauteil X?“). Das System sucht nun in einer speziellen Datenbank nicht nur nach dem Stichwort „Bauteil X“, sondern nach inhaltlich passenden Textabschnitten, die eine Antwort enthalten könnten.

Augmentation

Die gefundenen Textstellen werden als Kontext mit der ursprünglichen Frage verknüpft. Intern erhält die KI also die Anweisung: „Nutze die folgenden Hintergrundinformationen [Textauszug A, Textauszug B], um diese Frage zu beantworten: Wie hoch ist die Toleranz ...?“.

Generation

Das Sprachmodell generiert nun die Antwort. Dabei ist es durch den System-Prompt gezwungen, sich ausschließlich auf die gelieferten Informationen zu stützen und im Idealfall die Quelle („Siehe technisches Handbuch, Seite 42“) zu nennen.

2.4 VEKTOREN UND EMBEDDINGS

Damit der Computer verstehen kann, dass eine Suche nach „Rechnung“ auch Ergebnisse zu „Faktura“ oder „Zahlungsbeleg“ finden soll, nutzt RAG die sogenannte Semantische Suche.

Hierfür werden Texte in Vektoren (Embeddings) umgewandelt. Das sind lange Zahlenreihen, die die Bedeutung eines Wortes oder Satzes in einem mathematischen Raum repräsentieren. In diesem Raum liegen Begriffe mit ähnlicher Bedeutung (z. B. „Hund“ und „Katze“) räumlich näher beieinander als unähnliche Begriffe (z. B. „Hund“ und „Bohrmaschine“).

Eine Vektordatenbank speichert diese Zahlenreihen. Wenn ein Nutzender eine Frage stellt, wird auch diese Frage in einen Vektor verwandelt. Die Datenbank berechnet dann blitzschnell, welche gespeicherten Dokumenten-Vektoren der Frage mathematisch am nächsten kommen („Nearest Neighbor Search“). So findet das System relevante Antworten, selbst wenn der Nutzende ganz andere Wörter verwendet als im Dokument stehen, semantisch aber das gleiche bedeuten.

2.5 PRINZIP DES „CHUNKING“

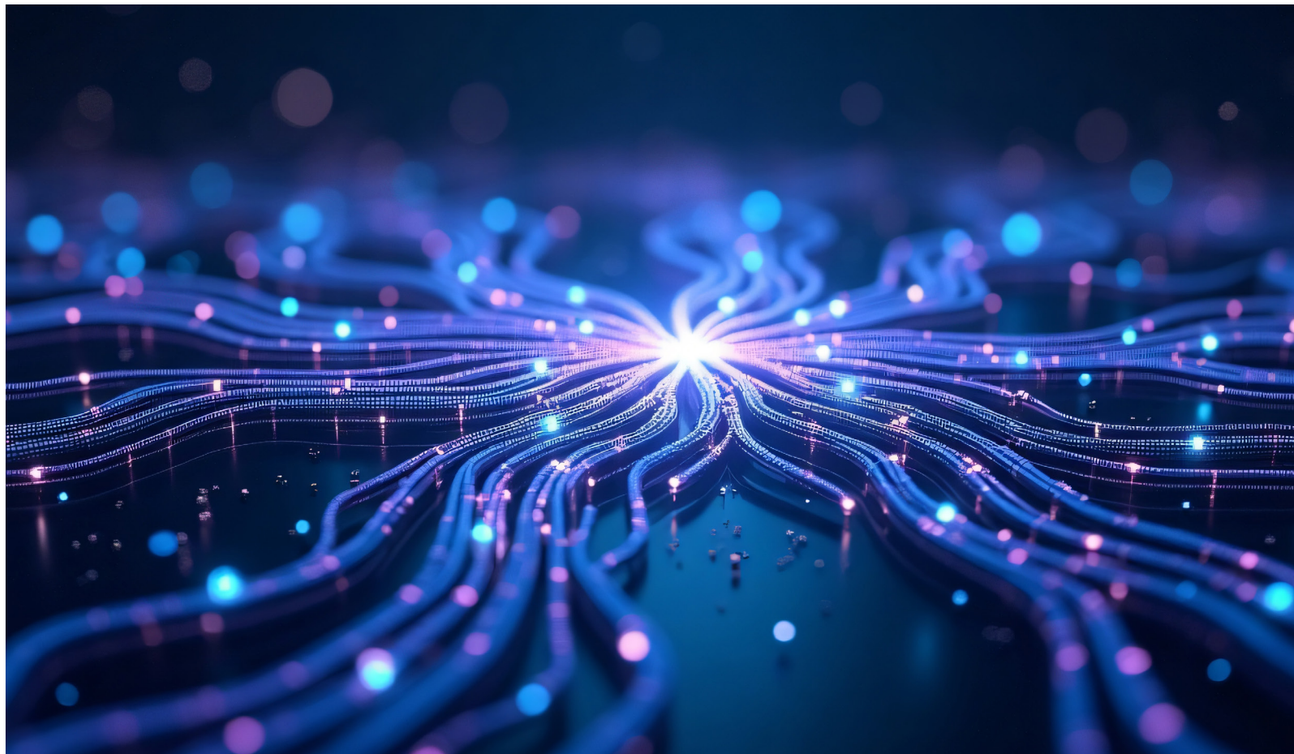
Ein KI-Modell kann nicht ein 500-seitiges PDF am Stück lesen und verstehen, um eine winzige Detailfrage zu beantworten. Daher müssen Daten vorab zerlegt werden. Dieser Prozess ist das sogenannte Chunking.

Dabei werden Dokumente in kleine, sinnvolle Häppchen (Chunks) unterteilt, z. B. Absätze oder logische Blöcke von 500 Zeichen. Wichtig ist hierbei oft eine „Überlappung“ (Overlap): Der erste Chunk enthält das Ende des vorherigen Absatzes, damit der Sinnzusammenhang an den Schnittstellen nicht verloren geht. Die Qualität dieses Chunkings entscheidet maßgeblich darüber, ob die KI später die präzise Antwort findet oder den Kontext übersieht.

2.6 KONTEXT-FENSTER

Warum braucht man RAG überhaupt, wenn moderne KIs immer größere Textmengen verarbeiten können? Die Erklärung hat mit dem Kontext-Fenster (Context Window) zu tun, welches jedes KI-Modell hat. Das ist eine Obergrenze an Informationen, die es gleichzeitig im „Arbeitsgedächtnis“ halten kann (gemessen in Tokens, also den Verarbeitungseinheiten eines Sprachmodells wie Wörter oder Satzzeichen).

Auch wenn Modelle großer Anbieter wie OpenAI, Google oder Anthropic große Fenster haben, ist es ineffizient und teuer, bei jeder Anfrage das gesamte Firmenwissen hochzuladen. Zudem sinkt die Präzision der Antworten, wenn das Modell in zu vielen irrelevanten Informationen „ertrinkt“ („Lost in the Middle“-Phänomen). RAG filtert die Informationen vorab, sodass nur die relevantesten 3-5 Textstücke im Kontext-Fenster landen. Dies spart Kosten und erhöht die Antwortqualität massiv.



3 BEDEUTUNG UND VORTEILE FÜR KMU

Für kleine und mittlere Unternehmen (KMU) ist Wissen oft die wichtigste Ressource und gleichzeitig die am schwersten zugängliche. Informationen liegen verstreut in PDF-Handbüchern, E-Mail-Archiven, Netzlaufwerken und den Köpfen langjähriger Mitarbeitender. RAG bietet hier nicht nur eine technische Spielerei, sondern einen echten wirtschaftlichen Hebel.

3.1 STRATEGISCHE WETTBEWERBSVORTEILE

In einem Markt, der immer schnellere Reaktionszeiten fordert, wird der sofortige Zugriff auf valides Wissen zum Wettbewerbsfaktor. KMU, die RAG einsetzen, können die Lücke zu Großkonzernen schließen, da sie ihre Flexibilität nun mit den analytischen Fähigkeiten großer KI-Modelle koppeln, ohne deren immense Entwicklungskosten tragen zu müssen.

Hebung des internen Datenschatzes

Viele KMU sitzen auf riesigen Mengen unstrukturierter Daten, die bisher „totes Kapital" waren. RAG macht dieses Wissen in Sekunden abfragbar und operationalisierbar, ohne dass es manuell in Datenbanken eingepflegt werden muss.

Demokratisierung von Expertenwissen

Spezialwissen (z. B. über alte Maschinenbaureihen oder komplexe Compliance-Regeln) ist oft bei wenigen Expert:innen konzentriert. RAG macht dieses Wissen für alle Abteilungen, vom Vertrieb bis zur Technik, verfügbar und verständlich. Das reduziert die Abhängigkeit von einzelnen Kopfmonopolen.

Skalierbarkeit ohne Personalaufbau

Ein RAG-System kann hunderte Kundenanfragen gleichzeitig beantworten oder Dutzende neue Mitarbeitende parallel bei der Einarbeitung unterstützen. Dies erlaubt Wachstum, ohne dass der administrative Overhead linear mitwächst.

3.2 KONKRETE VORTEILE GEGENÜBER ANDEREN KI-LÖSUNGEN

Warum sollten sich Mittelständler gerade für RAG entscheiden und nicht einfach öffentliche Chatbots wie ChatGPT, Gemini, Claude und Co. nutzen oder ein eigenes Modell trainieren? Die Antwort liegt in den folgenden Vorteilen:

Kosteneffizienz

Das Trainieren eines eigenen Sprachmodells kostet Hunderttausende Euro und erfordert spezialisierte Data Scientists. RAG nutzt hingegen günstige Standard-Modelle „von der Stange" und füttert sie nur mit den nötigen Infos. Die Implementierungskosten sind ein Bruchteil dessen, was ein Fine-Tuning kosten würde.

Absolute Aktualität

In dynamischen Branchen ändern sich Preise, Lagerbestände oder Vorschriften täglich. Ein klassisch trainiertes KI-Modell ist schon am Tag nach dem Training veraltet. Bei RAG tauschen Sie lediglich das zugrundeliegende Dokument (z. B. die Preisliste-PDF) aus, und die KI gibt ab der nächsten Sekunde die neuen Preise wieder. Es ist kein erneutes, teures Training nötig.

Datensouveränität & Datenschutz

Nutzen Mitarbeitende öffentliche Chatbots, landen Firmendaten oft auf US-Servern zum Training der KI. Bei einem eigenen RAG-System (insbesondere bei lokalen Installationen) haben Sie die volle Kontrolle. Das KI-Modell sieht immer nur den kleinen Textschnipsel, der für die aktuelle Frage nötig ist. Es lernt nicht dauerhaft aus Ihren Daten und speichert diese nicht ab.

Verlässlichkeit & Quellenangabe

Ein großes Problem generativer KI sind „Halluzinationen" (das Erfinden von Fakten). RAG zwingt die KI, Antworten nur auf Basis der gefundenen Dokumente zu geben. Zudem kann das System die Quelle direkt verlinken („Information gefunden in: Wartungshandbuch V2, Seite 14"). Dies schafft Vertrauen und ermöglicht eine schnelle Überprüfung durch den Menschen.

3.3 PRAXISBEISPIELE: ANWENDUNG IM MITTELSTAND

RAG ist keine abstrakte Zukunftsmusik, sondern löst bereits heute konkrete Probleme in verschiedenen Unternehmensbereichen:

Intelligenter Kundensupport (First-Level-Support)

Anstatt Kund:innen in Warteschleifen hängen zu lassen, kann ein RAG-Chatbot technische Fragen sofort beantworten. Er greift dabei auf tausende Seiten von Produkthandbüchern, FAQ-Listen und vergangenen Problemlösungen zu.
Beispiel: Ein Kunde fragt: „Welches Öl braucht die Maschine X-200 bei Minusgraden?" Der Bot findet die Info im PDF-Handbuch auf Seite 56 und antwortet präzise und rund um die Uhr.

Onboarding & interner Wissensmanager

Neue Mitarbeitende verbringen oft Wochen damit, Prozesse zu verstehen und Dokumente zu suchen. Ein interner „Firmen-GPT" fungiert als Mentor.
Beispiel: Ein neuer Vertriebler fragt: „Wie rechnen wir Reisekosten nach Österreich ab?" Das System zieht die Antwort aus der aktuellen Reiserichtlinie im Intranet und erklärt den Prozess schrittweise.

Unterstützung für Techniker & Montage

Servicetechniker:innen vor Ort müssen oft komplexe Fehler an Anlagen beheben. Statt ordnerweise Papier zu wälzen, nutzen sie eine RAG-App auf dem Tablet.
Beispiel: „Fehlercode E-404 bei Anlage B". Das System durchsucht alle Serviceberichte der letzten 10 Jahre nach diesem Fehler und schlägt die Lösung vor, die beim letzten Mal funktioniert hat.

Dokumentenanalyse & Ausschreibungen

KMU müssen oft umfangreiche Ausschreibungsunterlagen oder Verträge prüfen. RAG kann genutzt werden, um spezifische Risiken oder Anforderungen aus hunderten Seiten Text zu extrahieren.
Beispiel: „Liste alle Anforderungen an die IT-Sicherheit auf, die im Lastenheft des Kunden XY auf den Seiten 1-200 genannt werden." Das spart Stunden manueller Lesezeit.

4 NACHTEILE UND HERAUSFORDERUNGEN

Obwohl RAG als effizienter Weg gilt, KI in Unternehmen zu bringen, ist es keine Plug-and-Play-Lösung ohne Risiken. Die Implementierung verschiebt die Komplexität oft nur weg vom Training des Modells, hin zur Verwaltung der Daten und der Infrastruktur. Für KMU ergeben sich einige zentrale Hürden, die vor Projektstart bewertet werden müssen.

4.1 DATENQUALITÄT

Der Erfolg eines RAG-Systems steht und fällt mit der Qualität der gefütterten Daten – nach dem Prinzip „Garbage In, Garbage Out". Das Sprachmodell kann nur so intelligent antworten, wie die Dokumente es zulassen, die ihm zur Verfügung gestellt werden. In der Praxis zeigt sich das an folgenden Punkten:

Unstrukturierte Daten

In vielen KMU liegen Informationen in gescannten PDF-Dokumenten, handgeschriebenen Notizen oder chaotischen Excel-Tabellen vor. Ein RAG-System kann diese nicht ohne Weiteres lesen. Schlechte OCR (Optical Character Recognition, also Texterkennung) führt dazu, dass „Rechnung" als „Rechnun8" gelesen wird. Das Dokument wird dann bei einer Suche nicht gefunden.

Widersprüche & Veralterung

Wenn im System noch eine Preisliste von 2019 liegt, die der von 2024 widerspricht, kann die KI nicht entscheiden, welche wahr ist. Sie wird möglicherweise beide Informationen vermischen oder die falsche wählen. Datenbereinigung ist daher oft der zeitaufwendigste Teil des Projekts.

Kontext-Verlust durch Chunking

Werden Dokumente mechanisch zerteilt (z. B. strikt alle 500 Zeichen), können wichtige Zusammenhänge zerrissen werden. Eine Tabelle könnte in der Mitte geteilt werden, sodass die Spaltenüberschriften im ersten Teil landen und die Daten im zweiten. Für die KI werden die Zahlen im zweiten Teil damit nutzlos.

4.2 TECHNISCHE KOMPLEXITÄT UND FEHLENDE FACHKRÄFTE

RAG ist architektonisch komplexer als die bloße Nutzung eines Chatbots. Es erfordert das Zusammenspiel mehrerer Komponenten: Vektordatenbank, Embedding-Modell, LLM-API und Orchestrierungsschicht (siehe Kapitel 5). Die zentralen Herausforderungen liegen dabei vor allem in diesen Bereichen:

Wartungsaufwand

Diese Komponenten müssen gewartet und aktualisiert werden. Ändert sich die API (Schnittstelle) des LLM-Anbieters oder die Struktur der Datenbank, muss der Code angepasst werden.

Integration in Bestands-IT

Die größte Hürde ist oft nicht die KI selbst, sondern die Anbindung an Systeme wie Cloud-Speicher oder Netzlaufwerke. Hierfür werden spezialisierte Schnittstellen benötigt, die gepflegt werden müssen.

Fachkräftemangel

Für die Einrichtung wird Personal benötigt, das Verständnis von Data Engineering und KI-Architekturen hat. Solche Expert:innen sind auf dem Arbeitsmarkt aktuell schwer zu finden und teuer.

4.3 KOSTENSTRUKTUR UND SKALIERBARKEIT

Während RAG initial günstiger ist als das Trainieren eines eigenen Modells, entstehen im laufenden Betrieb variable Kosten, die schwer vorhersehbar sein können:

Token-Kosten

Bei jeder Anfrage bezahlt man nicht nur für die Antwort der KI, sondern auch für den „gelesenen" Kontext. Wenn ein Nutzender eine einfache Frage stellt, das System aber 10 Seiten Hintergrundinformationen (Kontext) an die KI schickt, steigen die Kosten pro Anfrage signifikant. Bei hohem Nutzeraufkommen skaliert dieser Kostenfaktor linear mit.

Infrastruktur

Vektordatenbanken benötigen Speicher und Rechenleistung. Hochperformante Cloud-Lösungen verursachen monatliche Fixkosten, die je nach Datenmenge steigen.

4.4 DATENSCHUTZ UND ZUGRIFFSKONTROLLE

Dies ist oft der kritischste Punkt für Unternehmen: Ein RAG-System zentralisiert Wissen, was Gefahren mit sich bringen kann. Das betrifft insbesondere folgende Risikofelder:

Interne Sicherheit

Wenn alle Unternehmensdokumente in einen großen Vektorspeicher geladen werden, könnte ein Praktikant den Chatbot fragen: „Wie viel verdient der Geschäftsführer?" Findet das System die Gehaltsliste im Speicher, wird es die Antwort liefern. RAG-Systeme müssen daher komplexe Berechtigungskonzepte (Access Control Lists) der Originalquellen spiegeln, was eine technisch anspruchsvolle Aufgabe ist.

Datenabfluss an Anbieter

Nutzen Sie z. B. Modelle von US-Anbietern, verlassen Fragmente Ihrer internen Dokumente (der Kontext) kurzzeitig Ihren Server zur Verarbeitung. Auch wenn Anbieter versichern, diese Daten nicht für das Training zu nutzen, stellt dies für sensible Branchen (Gesundheit, Finanzen) oft eine Compliance-Hürde dar.

5 VON DEN DATEN ZUM BETRIEB

Der Weg von einem Ordner voller PDF-Dateien zu einem intelligenten Chatbot, der Fragen zu diesen Dokumenten beantwortet, ist ein technischer Prozess, der Sorgfalt erfordert. Während die KI-Modelle selbst oft „Out-of-the-Box" funktionieren, liegt die eigentliche Arbeit im Aufbau der Datenpipeline (ETL) und der Feinjustierung der Suche.

5.1 TECHNISCHE VORAUSSETZUNGEN

Bevor die erste Zeile Code geschrieben wird, muss der Technologie-Stack definiert werden. Für KMU hat sich eine Architektur aus vier Kernkomponenten bewährt:

1. Das Sprachmodell (LLM)

Das Herzstück jeder RAG-Architektur ist das Sprachmodell (LLM). Es ist die Komponente, die den abgerufenen Kontext liest, versteht und die finale Antwort formuliert. Bei der Auswahl stehen Unternehmen vor einer fundamentalen Richtungsentscheidung: Nutzen wir die Rechenpower großer Anbieter in der Cloud oder betreiben wir die KI im eigenen Rechenzentrum?

Option A: Cloud-Lösungen (Software-as-a-Service, kurz SaaS / Managed Services)

Dies ist der Standardweg für 90 % der Unternehmen, die schnell starten wollen. Sie mieten die Intelligenz von Anbietern wie OpenAI, Anthropic oder Google per Schnittstelle (Application Programming Interface, kurz API). Sie erhalten sofortigen Zugriff auf die leistungsfähigsten Modelle (sogenannte „Frontier Models"). Diese Modelle verfügen über die besten schlussfolgernden Fähigkeiten und produzieren die qualitativ hochwertigsten Texte. Zudem entfällt jegliche Hardware-Investition. Sie zahlen variabel nach Verbrauch (Pay-per-Token).

Ein häufiger Kritikpunkt ist der Datenabfluss. Hier muss jedoch differenziert werden: Die Nutzung der öffentlichen Chatbot-Webseiten ist für sensible Daten oft tabu. Die Verwendung derselben Modelle über Enterprise-Partner kann jedoch einen völlig anderen Rechtsrahmen bieten. Hier können Sie z. B. Auftragsverarbeitungsverträge abschließen, die garantieren, dass Ihre Daten nicht zum Training der KI verwendet werden und die Server (je nach Konfiguration) innerhalb der EU stehen.

Die Kosten sind Betriebsausgaben. Das ist günstig für den Einstieg, skaliert jedoch linear. Wenn tausende Mitarbeitende das System täglich nutzen, können die monatlichen API-Rechnungen signifikant werden.

Option B: On-Premise / Lokal (Open Source)

Für Unternehmen mit hohen Anforderungen an die Datensouveränität (z. B. im Gesundheitswesen, Verteidigungssektor oder bei strengsten NDA-Vorgaben) ist der lokale Betrieb die Alternative. Hierbei laden Sie Modelle mit „offenen Gewichten" wie Llama 3 (Meta) oder Mistral (Mistral AI) auf Ihre eigenen Server herunter. Sie haben die Datenhoheit. Kein Byte an Daten verlässt Ihr Firmennetzwerk. Zudem sind Sie unabhängig von Preiserhöhungen oder Ausfällen eines Cloud-Anbieters.

Unterschätzen Sie allerdings nicht die Hardware-Anforderungen. Um ein Modell zu betreiben, das annähernd so gut ist wie die neusten Frontier-Modelle, benötigen Sie spezialisierte Server mit leistungsstarken Grafikkarten (GPUs) und viel VRAM (Videospeicher). Ein einziges professionelles KI-Setup kann schnell mittlere fünfstellige Hardware-Kosten verursachen. Sie benötigen außerdem intern IT-Personal, das sich mit „LLM Ops" auskennt. Modelle müssen quantisiert (komprimiert), aktualisiert und überwacht werden. Zudem hinken lokale Modelle in ihrer „Intelligenz" den Frontier-Modellen oft noch leicht hinterher, auch wenn die Lücke kleiner wird.

2. Die Vektordatenbank

Klassische Datenbanken (wie SQL) sind darauf ausgelegt, exakte Treffer zu finden (z. B. „Zeige Kunde Nr. 123"). RAG erfordert jedoch eine Datenbank, die semantische Ähnlichkeiten versteht. Hier kommen Vektordatenbanken ins Spiel. Sie speichern Daten nicht in Tabellen, sondern in einem hochdimensionalen mathematischen Koordinatensystem. Wenn Sie eine Frage stellen, sucht die Datenbank nicht nach dem Wortlaut, sondern berechnet, welche gespeicherten Dokumente der Frage im Raum am nächsten liegen. Das ermöglicht es, Antworten zu finden, selbst wenn kein einziges Wort der Frage im Originaltext vorkommt.

Unternehmen müssen sich entscheiden: Nutzen Sie eine native Vektordatenbank (z. B. Pinecone, Weaviate, Qdrant), die speziell für diese Aufgabe gebaut wurde und performant skaliert? Oder nutzen Sie eine Erweiterung bestehender Systeme (z. B. PostgreSQL mit pgvector oder Elasticsearch)? Für viele Mittelständler ist die PostgreSQL-Variante attraktiv, da IT-Abteilungen das Know-how und die Infrastruktur dafür oft schon im Haus haben. Native Cloud-Lösungen punkten hingegen mit Wartungsfreiheit und Geschwindigkeit bei Millionen von Datensätzen.

Eine gute Vektordatenbank muss außerdem „Hybrid Filtering" beherrschen. Das bedeutet: Sie wollen oft nicht nur „inhaltlich ähnliche" Dokumente finden, sondern die Suche einschränken (z. B. „Suche nach Problemlösungen für Maschine A, aber nur in Dokumenten, die nach 2022 erstellt wurden"). Die Datenbank muss also die semantische Unschärfe der KI mit der harten Logik klassischer Filter verknüpfen.

3. Das Embedding-Modell

Dies ist die unscheinbarste, aber vielleicht wichtigste Komponente. Bevor Text in die Datenbank gelangt oder eine Frage verarbeitet wird, muss dieses Modell die menschliche Sprache in eine lange Zahlenreihe (Vektor) verwandeln. Die Qualität dieses Modells bestimmt, wie „klug" das System Zusammenhänge erkennt. Ein gutes Embedding-Modell unterscheidet Bedeutung im Kontext. Es weiß, dass das Wort „Bank" in „Ich sitze auf der Bank" mathematisch weit weg liegt von „Ich überweise Geld an die Bank". Schlechte Modelle vermischen diese Bedeutungen, was zu irrelevanten Suchergebnissen führt.

Für exportorientierte KMU ist entscheidend, ob das Modell sprachenübergreifend funktioniert. Moderne Modelle erlauben es, eine Frage auf Deutsch zu stellen („Wie wechsle ich die Sicherung?"), finden die Antwort aber auch dann, wenn sie nur in einem englischen oder spanischen Handbuch steht. Der Vektor für „Sicherung" (DE) und „Fuse" (EN) ist in diesen Modellen fast identisch.

Zur Auswahl eines passenden Modells orientieren Sie sich z. B. am sogenannten MTEB-Leaderboard (Massive Text Embedding Benchmark), aber testen Sie mit eigenen Daten. Cloud-Embeddings sind einfach einzubinden, aber Ihre Daten verlassen kurzzeitig das Haus. Lokale Open-Source-Modelle laufen auf Ihrer eigenen CPU/GPU, sind datenschutzrechtlich unbedenklich und oft spezialisierter anpassbar, erfordern aber mehr Integrationsaufwand.

4. Orchestrierungs-Frameworks

Ohne ein solches Framework müssten Entwickler hunderte Zeilen komplexen Code schreiben, um API-Aufrufe zu verwalten, Datenformate zu konvertieren und Fehler abzufangen. Orchestrierungs-Tools standardisieren diese Prozesse. LangChain ist das derzeit bekannteste Framework auf dem Markt. Es funktioniert wie ein Baukasten aus Lego-Steinen. Es ermöglicht Entwicklern, verschiedene Komponenten modular zu verknüpfen („Chaining“).

Der große Vorteil für KMU: LangChain ist modell-agnostisch, also nicht an ein bestimmtes Modell gebunden. Wenn Sie heute GPT-4 nutzen und morgen auf ein günstigeres Modell oder eine lokale Open-Source-Alternative wechseln wollen, müssen Sie oft nur eine einzige Zeile im Code ändern, statt die gesamte Anwendung neu zu schreiben.

Während LangChain sehr breit aufgestellt ist, spezialisiert sich LlamaIndex (früher GPT Index) extrem auf die effiziente Verarbeitung und Indizierung von Daten. Wenn Ihr Fokus darauf liegt, sehr komplexe, strukturierte Datenquellen (z. B. verschachtelte SQL-Datenbanken oder riesige Excel-Reports) durchsuchbar zu machen, bietet LlamaIndex oft die besseren, spezialisierten Werkzeuge für das Retrieval.

Ein entscheidender Mehrwert dieser Frameworks ist das Management des „Gesprächsgedächtnisses“ (Memory). Sie sorgen dafür, dass der Chatbot noch weiß, was vor drei Nachrichten gesagt wurde. Zudem ermöglichen sie den Bau von „Agenten“: Das sind KI-Systeme, die nicht nur antworten, sondern basierend auf der Antwort Aktionen auslösen können (z. B. „Suche die Info im Handbuch UND schreibe danach eine E-Mail an den Support“).

5.2 ETL-PROZESS: DATENAUFBEREITUNG

Der wichtigste Erfolgsfaktor eines RAG-Systems ist die Qualität der Datenbasis. Das Prinzip „Garbage In, Garbage Out“ gilt hier uneingeschränkt. Der Prozess der Datenaufbereitung (Extract, Transform, Load) gliedert sich in vier entscheidende Schritte:

Schritt 1: Extraktion & Bereinigung

Zunächst müssen Texte aus den Quellsystemen (SharePoint, Netzlaufwerke, Wikis) extrahiert werden. Die Herausforderung liegt oft in der Formatierung:

» **PDF-Problematik**
Kopf- und Fußzeilen, Seitenzahlen oder Spaltenlayouts in PDFs verwirren die KI. Diese müssen algorithmisch entfernt werden, damit der Fließtext Sinn ergibt.

» **Tabellen**
Komplexe Tabellen in Dokumenten gehen beim reinen Text-Extrakt oft verloren. Hier sind spezielle Parser notwendig, die Tabellenstrukturen erkennen und in Text beschreiben (z. B. „Zeile 1: Umsatz 2023, Spalte B: 50.000 €“).

Schritt 2: Chunking

Da KI-Modelle nur eine begrenzte Menge Text auf einmal verarbeiten können, müssen lange Dokumente in kleine Häppchen („Chunks“) zerlegt werden. Die Strategie hierbei ist entscheidend:

» **Fixed-Size Chunking**
Der Text wird strikt nach z. B. 500 Zeichen getrennt. Das ist einfach, zerreißt aber oft Sätze oder Sinnzusammenhänge.

» **Semantic Chunking**
Intelligente Algorithmen trennen den Text dort, wo ein Thema endet und ein neues beginnt (z. B. nach Absätzen oder Überschriften).

» **Der Overlap (Überlappung)**
Um Kontextverlust an den Schnittstellen zu vermeiden, sollten sich Chunks überlappen (z. B. 10-20 %). So weiß der zweite Chunk noch, worum es am Ende des ersten ging.

Schritt 3: Embedding & Metadaten

Jeder Text-Chunk wird nun in einen Vektor umgewandelt. Wichtig ist hierbei die Anreicherung mit Metadaten. Speichern Sie zu jedem Vektor Informationen wie „Erstellungsdatum“, „Autor“ oder „Dokumententyp“. Dies ermöglicht später Filterungen wie: „Suche nur in Dokumenten, die neuer als 2022 sind.“

Schritt 4: Indexierung

Die Vektoren und Metadaten werden in die Vektordatenbank geladen und indexiert. Ein guter Index sorgt dafür, dass auch in Millionen von Dokumenten die passende Antwort in wenigen Millisekunden gefunden wird.

5.3 SCHRITT FÜR SCHRITT ZUR RAG-ANWENDUNG

Wie läuft ein Entwicklungsprojekt in der Praxis ab? Ein iteratives Vorgehen hat sich bewährt:

Phase 1: Prototyping (Minimum Viable Product, kurz MVP)

Starten Sie klein. Wählen Sie einen klar abgegrenzten Anwendungsfall (z. B. „Fragen zur IT-Sicherheitsrichtlinie“) und laden Sie 10 bis 50 Dokumente manuell in ein Framework. Ziel ist es, schnell ein Gefühl dafür zu bekommen, wie gut die KI auf Ihre spezifischen Daten reagiert.

Phase 2: Optimierung des Retrievals (Hybrid Search & Re-Ranking)

Oft stellt man fest: Die KI antwortet gut, wenn sie die richtige Info hat, aber sie findet die Info nicht immer. Hier helfen fortgeschrittene Techniken:

» **Hybrid Search**
Kombinieren Sie die Vektorsuche (findet Konzepte) mit der klassischen Stichwortsuche (findet exakte Begriffe wie Artikelnummern „X-200“). Vektorsuche allein scheitert oft an spezifischen Produktcodes.

» **Re-Ranking**
Fügen Sie einen zweiten Prüfschritt ein. Ein spezialisiertes KI-Modell (Cross-Encoder) schaut sich die Top-10 Ergebnisse der Datenbank noch einmal genau an und sortiert sie neu nach Relevanz, bevor sie an das Sprachmodell gehen. Dies kann die Trefferquote erhöhen.

Phase 3: Evaluation

Bevor Sie live gehen, müssen Sie messen, ob das System verlässlich ist. Bauen Sie ein „Golden Dataset“ auf: Eine Liste von 50 typischen Fragen und den dazugehörigen idealen Antworten (von Expert:innen geschrieben). Prüfen Sie anschließend:

» **Context Recall**
Wurde die richtige Textstelle überhaupt gefunden?

» **Faithfulness**
Hält sich die Antwort der KI an die Fakten im Text oder wurde etwas Erfundenes hinzugefügt?

Phase 4: Integration & Feedback-Loop

Integrieren Sie den Bot dort, wo die Nutzenden arbeiten (Microsoft Teams, Intranet, CRM). Bauen Sie zwingend Buttons für „Daumen hoch / Daumen runter“ unter jede Antwort ein. Dieses Feedback ist besonders wertvoll, um schlechte Antworten zu identifizieren, fehlende Dokumente zu ergänzen und das System kontinuierlich zu verbessern.

6 FAZIT

Retrieval Augmented Generation (RAG) ist weit mehr als nur ein aktueller Technologie-Trend oder ein kurzlebiges Schlagwort in der IT-Welt. Es markiert einen fundamentalen Wandel im betrieblichen Wissensmanagement. Für kleine und mittlere Unternehmen (KMU) stellt diese Technologie den entscheidenden Schlüssel dar, um die mächtigen Fähigkeiten moderner Künstlicher Intelligenz überhaupt erst produktiv und sicher nutzbar zu machen. Ohne RAG bleibt KI ein eloquenter, aber unwissender Gesprächspartner. Mit RAG wird sie zum kompetenten Fach-Assistenten, der Ihr Unternehmen im Detail kennt.

Die Vorteile liegen auf der Hand: Die Technologie ermöglicht es, den oft brachliegenden Schatz an internen Daten zu heben und sekundenschnell operationalisierbar zu machen. Dies steigert nicht nur die Effizienz im Tagesgeschäft, indem Suchzeiten drastisch reduziert werden, sondern demokratisiert auch Expertenwissen. Wenn ein Junior-Mitarbeiter im Kundenservice mit der gleichen Präzision antworten kann wie ein Senior-Ingenieur mit 20 Jahren Erfahrung, weil das System ihm die korrekten Informationen liefert, stärkt das die gesamte Organisation und macht sie resilienter gegen den Fachkräftemangel.

Doch der Weg dorthin ist kein reines IT-Projekt, sondern eine strategische Aufgabe. Die Einführung von RAG zwingt Unternehmen dazu, sich ehrlich mit ihrer eigenen Datenqualität auseinanderzusetzen. Die Erkenntnis „Garbage In, Garbage Out“ wird hier zur unmittelbaren Realität: Ein RAG-System kann nur so gut sein wie die Dokumente, die es liest. Damit wird die Datenpflege von einer lästigen Pflichtaufgabe zu einem direkten Wettbewerbsfaktor. Wer heute beginnt, seine Datenstrukturen zu bereinigen und digitale Silos aufzubrechen, schafft das Fundament für die Wettbewerbsfähigkeit von morgen.

Der Blick in die Zukunft zeigt, dass diese Entwicklung erst am Anfang steht. Während aktuelle Systeme primär Text verarbeiten, entwickeln sich RAG-Anwendungen rasant in Richtung Multimodalität weiter. Sie werden zukünftig technische Zeichnungen interpretieren oder komplexe Excel-Tabellen analysieren können. Zudem wird sich die Technologie von reinen Frage-Antwort-Systemen hin zu „Agenten-Systemen“ wandeln, die basierend auf den gefundenen Informationen auch Handlungen vorschlagen oder Prozesse im ERP-System anstoßen können.

Für den Mittelstand lautet die Empfehlung daher: Warten Sie nicht auf die „perfekte“ All-in-One-Lösung. Die Technologie ist reif genug für den produktiven Einsatz. Starten Sie mit kleinen, überschaubaren Pilotprojekten. Etwa mit einem internen Chatbot für HR-Fragen oder einer Suchhilfe für die technische Dokumentation. Sammeln Sie Erfahrungen mit der Aufbereitung Ihrer Daten und der Akzeptanz durch die Mitarbeitenden. Denn der größte Wert liegt nicht in der Software selbst, sondern in der Lernkurve, die Ihr Unternehmen durchläuft. Die Zeit, das eigene Unternehmenswissen intelligent zu nutzen, ist jetzt.

Das Mittelstand-Digital Zentrum Augsburg gehört zu Mittelstand-Digital. Mit dem Mittelstand-Digital Netzwerk unterstützt das Bundesministerium für Wirtschaft und Energie die Digitalisierung in kleinen und mittleren Unternehmen und dem Handwerk.

Das Mittelstand-Digital Netzwerk bietet mit den Mittelstand-Digital Zentren und der Initiative IT-Sicherheit in der Wirtschaft umfassende Unterstützung bei der Digitalisierung mit dem Schwerpunkt Künstliche Intelligenz. Kleine und mittlere Unternehmen profitieren von konkreten Praxisbeispielen und passgenauen, anbieterneutralen Angeboten zur Qualifikation und IT-Sicherheit. Das Bundesministerium für Wirtschaft und Energie ermöglicht die kostenfreie Nutzung der Angebote von Mittelstand-Digital.

Weitere Informationen finden Sie unter www.mittelstand-digital.de