Fraunhofer

Harriet Kasper | Dennis Klau

Unternehmensdokumente erschließen mit RAG

Wie mit Retrieval-Augmented Generation der Chat mit den eigenen Daten gelingt

Hrsg.: Katharina Hölzle, Oliver Riedel, Wilhelm Bauer, Thomas Renner, Matthias Peissner

Im Rahmen des

KI Fortschrittszentrum
LERNENDE SYSTEME UND KOGNITIVE ROBOTIK

Vorwort





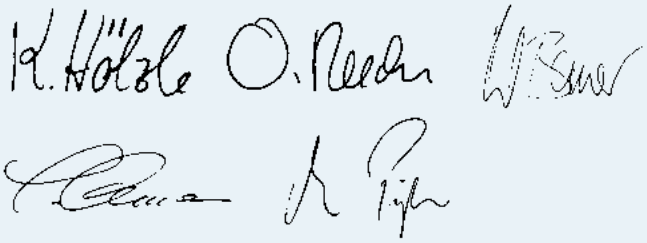
Künstliche Intelligenz (KI) ist eine der zentralen Technologien für die Zukunft. Ihre Einführung und der Einsatz fordern Unternehmen im besonderen Maß heraus. Es gilt, das Potenzial zu erkennen und dieses wirtschaftlich nutzbar zu machen. Lassen Sie sich dabei von Europas größter Forschungskooperation auf dem Gebiet der KI, Cyber Valley, begleiten.

Mit dem KI-Fortschrittszentrum vom Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO und Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA und als Teil von S-TEC, dem Stuttgarter Technologie- und Innovationscampus, unterstützen wir Unternehmen dabei, das Potenzial von KI nutzbringend einzusetzen. An der Schnittstelle zwischen anwendungsorientierter Wissenschaft und exzellenter Forschung des Cyber-Valley-Konsortiums entwickeln wir innovative KI-Anwendungen für die Praxis und treiben damit die Kommerzialisierung von KI voran. Erklärtes Ziel ist dabei, menschzentrierte KI-Lösungen zu entwickeln. Denn nur wenn Menschen mit einer neuen Technologie intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann ihr Potenzial optimal ausgeschöpft werden.

Die Studienreihe »Lernende Systeme und Kognitive Robotik« des KI-Fortschrittszentrums gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Dabei wurden bereits übergreifende Themen wie Zuverlässigkeit, Erklärbarkeit (xAI), cloudbasierte Plattformen, Technologien und Einführungsstrategien sowie einzelne Anwendungsbereiche vorgestellt. Mit der Fortsetzung der Reihe kommen zahlreiche neue Themen hinzu. Diese reichen von der Bildverarbeitung über die effiziente Software- und Systemintegration in der Robotik, die Gestaltung von KI-Systemen, Sprachmodelle in der Praxis bis hin zu KI-basierten Assistenzsystemen und dem Beitrag von KI zur Fachkräftesicherung.

Diese Studie bietet einen kompakten Überblick über Retrieval-Augmented Generation (RAG), einen Ansatz, bei dem Informationen aus einer Auswahl von Dokumenten eines Unternehmens mithilfe großer Sprachmodelle erschlossen werden. Der aktuell häufigste RAG-Anwendungsfall sind spezifische Chatbots, die Mitarbeitende bei internen Aufgaben unterstützen. Im KI Fortschrittszentrum hat das Fraunhofer IAO mehrere RAG-Anwendungsfälle prototypisch umgesetzt; sie illustrieren in dieser Publikation die Bandbreite möglicher Lösungen. Des Weiteren wurden leitfadengestützte Interviews geführt, um Vorgehen und Herausforderungen bei der Umsetzung von RAG Produktivsystemen in großen Organisationen zu analysieren. Aus den Ergebnissen werden praxisnahe Handlungsempfehlungen abgeleitet, die Unternehmen dabei unterstützen, »den Chat mit den eigenen Daten« erfolgreich und verantwortungsvoll zu implementieren.

Wir wünschen Ihnen eine spannende Lektüre, und freuen uns, wenn wir in Zukunft auch Sie mit unserer Expertise auf Ihrem Weg zur menschzentrierten KI unterstützen dürfen.



Katharina Hölzle, Oliver Riedel, Wilhelm Bauer,
Thomas Renner, Matthias Peissner
Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO

v.l.n.r.: Univ.-Prof. Dr. habil. Katharina Hölzle, Univ.-Prof. Dr.-Ing. Oliver Riedel, MBA, Prof. Dr.-Ing. Wilhelm Bauer, Dipl.-Ing. Thomas Renner, Dr.-Ing. Matthias Peissner

Inhalt

Management Summary

Seit der Einführung von ChatGPT Ende 2022 hat sich der Einsatz generativer KI rasant verbreitet, wobei insbesondere interne Anwendungen und Chatbots, die auf Unternehmensdokumente zugreifen, zunehmend an Bedeutung gewinnen. Diese Systeme ermöglichen es Mitarbeitenden, spezifische, meist interne Informationen zu erschließen, ohne detaillierte Kenntnisse über die Informationsstruktur innerhalb der Dokumente und die Dokumentenablage zu benötigen. Die Studie bietet einen kompakten Überblick zu Retrieval-Augmented Generation (RAG), einem aktuellen Ansatz, um »den Chat mit den eigenen Daten« zu realisieren.

RAG koppelt ein generatives Sprachmodell (LLM) mit einem Retrieval-Modul, das relevante Informationen aus bereitgestellten Datenbeständen identifiziert und als Kontext an das LLM übergibt. Der Ansatz stellt einen Mittelweg dar: flexibler und wirtschaftlicher als Fine-Tuning eines LLMs, skalierbarer und nachvollziehbarer als reines Prompt Engineering, weil Quellen verlinkt und passagengenau geprüft werden können. Zugleich bleibt RAG kein Allheilmittel: Halluzinationen sind nicht vollständig vermeidbar und die Qualität der Ergebnisse hängt maßgeblich von Verfügbarkeit, Pflege und Governance ab.

Die Studie verwendet drei Quellentypen: erstens aktuelle Literatur, um RAG zu erklären. Zweitens Erfahrungswissen des Fraunhofer IAO aus prototypischen Umsetzungen in mehreren konkreten Anwendungsfällen zur Veranschaulichung. Drittens die Auswertung von fünf leitfadengestützten Interviews mit Personen, die an der Umsetzung von RAG-Produktivsystemen in großen Organisationen beteiligt waren bzw. sind, um die Praxisperspektive einzubeziehen.

Aus den Interviews leitet die Studie folgende zentrale Erkenntnisse ab:

- Interne, spezifische Chatbots sind aktuell die wichtigsten RAG-Produktivsysteme.
- Ihre Umsetzung erfolgt typischerweise inhouse und dauert rund ein Jahr. Häufig wurden im Zuge des ersten RAG-Projekts neue Teams zusammengestellt.
- Auch in den untersuchten größeren Organisationen sind RAG-Chatbots erst seit 2025 im Einsatz und werden stetig erweitert.
- Daten sind ein Engpass: Auswahl und Bereinigung der Dokumente durch den Fachbereich bestimmen Qualität und Aufwand. Bislang dominieren textbasierte Dokumente.
- Cloud-first-Strategien sind Standard, die Umsetzung erfolgt häufig mit Unterstützung des Plattformanbieters; Souveränität und mögliche Vendor-Lock-ins müssen in diesem Zusammenhang aktiv bewertet werden.
- Die Akzeptanz von RAG-Chatbots ist hoch, weil in deren Ausgabe Quellen verlinkt und relevante Passagen direkt einsehbar sind; die Anzeige eines Haftungsausschluss-Hinweises (Disclaimer) gehört zu den Good Practices.
- Die Messung von Effizienzgewinnen durch RAG-Chatbots ist herausfordernd; Nutzungsmetriken allein genügen nicht.
- Trotzdem ist die Nachfrage hoch und Unternehmen setzen aktuell eine »Fertigungsstraße« für weitere interne RAG-Chatbots um.

Zusammen mit den Erfahrungen aus den Prototypen wurden in Workshops durch das Fraunhofer IAO unter anderem folgende Handlungsempfehlungen erarbeitet:

- Mitarbeitende müssen die Möglichkeiten, Herausforderungen und Grenzen von KI kennenlernen. Schaffen Sie Berührungspunkte und schulen Sie Ihre Mitarbeitenden.
- Starten Sie klein und intern, mit einem spezifischen RAG-Chatbot als Machbarkeitsnachweis. Nutzen Sie dabei schon vorhandene Daten.
- Verlinken Sie Quellen konsequent und direkt ersichtlich.
- Sichern Sie für das RAG-Projekt Ressourcen in der IT (2–3 Entwickler) und im Fachbereich (1–2 Domänen-Expertinnen/-Experten).
- Schaffen Sie früh eine technische Blaupause und organisatorische Standards für weitere, ähnliche Anwendungsfälle.
- Führen Sie einen Disclaimer ein und nehmen Sie eine Risikoabschätzung für Fehlantworten vor.
- Evaluieren Sie das RAG-System kontinuierlich – zunächst per Stichproben mit Gold-Standard-Datensätzen, später ergänzend mit automatisierten Ansätzen wie LLM-as-a-Judge.
- Planen Sie Betrieb, Kostenkontrolle und Erweiterungen.
- Prüfen Sie Souveränität regelmäßig (Abhängigkeiten, Wechseloptionen, Kostenpfade).
- Stärken Sie übergreifend Datenstrategie und Data-Governance.

Kapitel 5 erläutert die Handlungsempfehlungen und klassifiziert sie in allgemeine Empfehlungen bezüglich KI, Empfehlungen für das erste RAG-Projekt, Empfehlungen für den Betrieb von RAG-Systemen und Zukunftsthemen.

Agentische KI und Agentic RAG markieren den nächsten Schritt zu autonomeren, toolgestützten Workflows – mit höherer Leistungsfähigkeit, aber auch mehr Komplexität und Kontrollbedarf.

Außerdem ist zu beachten, dass die wirtschaftliche Bedeutung von RAG-Systemen mit steigender Nutzung wächst: Betriebskosten (LLM-Aufrufe, Vorverarbeitung, OCR) und Personal für Pflege und Überwachung nehmen zu; daraus resultiert ein strategischer Vorteil für Open-Weight-Modelle und souveräne Architekturen.

Fazit der Studie ist: Zukunftsfähig sind nicht jene Unternehmen, die heute die meisten oder besten RAG-Anwendungsfälle umgesetzt haben, sondern jene, die durch die Umsetzung einer Datenstrategie ihren gesamten Datenbestand KI-ready machen.

Das Team Angewandte Künstliche Intelligenz im Forschungsbereich Digital Business am Fraunhofer IAO unterstützt Unternehmen unter anderem bei der Konzeption und prototypischen Umsetzung von RAG-Systemen, bei der Bewertung und Etablierung von KI-Datenstrategien und bei dem Thema Souveränität von KI-Lösungen. Es entwirft und evaluiert KI-Lösungsansätze für die Funktionsbereiche Kundenservice und Vertrieb sowie für die Automatisierung von Arbeitsabläufen in der Versicherungswirtschaft. Weitere Informationen dazu unter www.digital.iao.fraunhofer.de

1. Einleitung

Das erste Kapitel gibt einen Überblick über Hintergrund und Motivation der Studie, beschreibt die Methodik und den Aufbau der folgenden Kapitel.

1.1. Hintergrund und Motivation

ChatGPT konnte bereits 5 Tage nach seiner Vorstellung im November 2022 1 Million Nutzende verzeichnen. Die Webapplikation zeigt für jeden zugänglich, anschaulich und beeindruckend, wie sprachlich ausgefeilte Antwort-Texte auf unterschiedlichste Fragen in kürzester Zeit automatisch erstellt werden. Damit wurde ein anhaltender Hype um Generative Künstliche Intelligenz ausgelöst. ChatGPT von OpenAI verzeichnet nach eigenen Angaben Ende 2025 800 Millionen wöchentliche Nutzende[1] und liegt damit weit vor den Sprachmodell-Konkurrenten wie Meta AI, Gemini von Google, Perplexity oder Claude von Anthropic. Pro Tag werden mehrere Milliarden Prompts verarbeitet. Die repräsentative ARD/ZDF-Medienstudie hat 2025 erstmals die Nutzung von KI-Tools erhoben (ARD/ZDF-Forschungskommission, 2025): 22 Prozent vor allem jüngere Deutsche nutzen KI-Tools mindestens einmal pro Woche. Die KI-Zusammenfassung von Google verändert seit ihrer Einführung (05/2024 in den USA und 03/2025 in Deutschland) das Internet – Chatbots sind für viele eine neue Art und Weise, auf Wissen zuzugreifen.

Unternehmenslenker und Mitarbeitende wurden durch diese Entwicklungen inspiriert, individuell relevante Anwendungsfälle für Generative KI zu finden und zu erproben (Kintz et al., 2024). Hierzu besteht noch großer Informationsbedarf, den diese Studie adressiert. Betrachtet werden dabei **vor allem interne Chatbots mit spezifischem Wissen,** die als Assistenten Mitarbeitende bei ihren Aufgaben unterstützten. Das Thema KI in der Kundenkommunikation, also Chatbots, die mit Kundinnen und Kunden kommunizieren, wird in der Studie *Einsatz von KI mit Fokus Kundenkommunikation* (Drawehn & Pohl, 2023) des KI-Fortschrittszentrums behandelt.

Fokus dieser Studie ist es, **Retrieval-Augmented Generation (RAG) als aktuellen Ansatz für Chatbots mit Zugriff auf Unternehmensdokumente** zu erläutern sowie Erfahrungen und Handlungsempfehlungen in diesem Zusammenhang aufzuzeigen. Abbildung 1 illustriert einen einfachen Anwendungsfall, der die neue veränderte Arbeitsweise durch solche spezifischen internen Chatbots zeigt: Es ist nicht mehr notwendig zu wissen, wo Daten abgelegt sind, und auch eine Verknüpfung von Daten (Rechnung und Zahlungseingang) erfolgt automatisch.

1 https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/?utm_source=chatgpt.com

Abbildung 1: Sachbearbeiter chattet mit Unternehmensdokumenten. KI-generiertes Bild, erstellt mit DALL·E 3 (OpenAI).

RAG ist eine Kombination von zwei Technologien: einem generativen Sprachmodell (Generator) und einem Retrieval-Modul, das man sich als Suche oder Index von spezifischen Informationen vorstellen kann. Durch RAG verarbeitet ein Sprachmodell zusätzliche Informationen, z. B. interne Unternehmensdokumente, und basiert seine Antworten vor allem auf deren Inhalten. Eine weitere Möglichkeit, spezifisches Wissen durch einen Chatbot zur Verfügung zu stellen, ist es, ein Sprachmodell mit diesen zusätzlichen Informationen zu trainieren. Der RAG-Ansatz ist in den meisten Fällen derzeit wirtschaftlicher und weithin verbreitet und wird deshalb in dieser Studie vertieft.

Viele, vor allem große Organisationen, stellen ihren Mitarbeitenden einen internen Alltags-Chatbot zur Verfügung – zum Beispiel wurde bei der Fraunhofer-Gesellschaft (FhG) im Juni 2023 FhGenie eingeführt[2] (Weber et al., 2024). In der Regel handelt es sich dabei erstmal nur um den Zugang zu einem großen Sprachmodell (LLM), der den Anforderungen an Datenschutz, Vertraulichkeit und Informationssicherheit genügt. Damit können sich die Mitarbeitenden vertrauliche Dokumente zusammenfassen lassen oder Unterstützung bei der Formulierung von Texten erhalten. Spezifische, nicht öffentlich verfügbare Informationen, zum Beispiel zu Prozessen oder Regeln der Organisation, hat solch ein interner Alltags-Chatbot nicht. Außerdem stößt man mit großen Sprachmodellen auf Herausforderungen wie Halluzinationen[3], veraltetes Wissen oder nicht nachvollziehbare Schlussfolgerungsprozesse.

2 https://www.fraunhofer.de/de/presse/presseinformationen/2023/august-2023/fhgenie-fraunhofer-gesellschaft-fuehrt-internen-ki-chatbot-ein.html

3 Randnotiz: Als Halluzination bezeichnet man die Situation, in der ein LLM zwar sprachlich stimmigen und plausibel klingenden Output erzeugt, der allerdings faktisch falsch ist. Verschiedene Forscher plädieren für eine Umbenennung dieses Begriffs, da er das zugrunde liegende Problem nicht richtig beschreibt. Allerdings hat sich der Terminus bereits in Literatur und Sprachgebrauch etabliert.

Um dem zu begegnen, können in einem zweiten Schritt interne Datenquellen über einen RAG-Ansatz angebunden werden. Meist sind aber interne Anwendungen eines LLM wie ChatGPT abzugrenzen von Chatbots für spezifische Anwendungsfälle. Abbildung 1 illustriert zum Beispiel, wie ein Sachbearbeiter natürlichsprachlich nach den offenen Rechnungen fragt, statt die Information in unterschiedlichen Systemen (Rechnungsablage, Zahlungseingängen) zu suchen und zusammenzustellen. Solche spezifischen Anwendungsfälle, die im Unternehmen messbare Effizienzgewinne bringen, werden heute eher unabhängig von dem allgemeinen Unternehmens-Chatbot (Alltags-Chatbot) realisiert. Sie stehen zumeist einem ausgewählten Nutzendenkreis zur Verfügung und bieten in ihren Antworten direkt Links auf die verwendeten Quellen und Dokumente.

Interne Chatbots für spezifische Anwendungsfälle sind eine neue Art, Wissen zugänglich zu machen, und sind sehr gefragt. Wurden solche Chatbots mithilfe von erfahrenen Mitarbeitenden realisiert und getestet, benötigen die Mitarbeitenden, die sie bedienen, weniger Erfahrungswissen. Doch auch wenn Fachkräftemangel und hohe Fluktuation von Mitarbeitenden für ein Unternehmen keine Herausforderungen darstellen, so spricht in vielen Fällen, neben der Effizienzsteigerung, die zunehmende Möglichkeit oder auch Notwendigkeit, weitere Daten zu berücksichtigen, für den Einsatz von solchen Chatbots bzw. RAG-Systemen im Allgemeinen.

1.2. Überblick über die Studie

1.2.1. Zielsetzung und Zielgruppe

Zielsetzung dieser Studie ist es, wissenschaftlich fundiertes Orientierungswissen für die Einführung und den Betrieb von RAG-Systemen (insbesondere RAG-Chatbots) bereitzustellen. Die Inhalte richten sich dabei sowohl an Unternehmen, die RAG-Systeme bereits im Einsatz haben, als auch an Personen, welche die Möglichkeiten und Funktionsweisen des RAG-Ansatzes verstehen und für ihr Unternehmen einschätzen möchten. Erfahrenen Nutzern werden sowohl technische Weiterentwicklungsmöglichkeiten aufgezeigt als auch Anwendungswissen vermittelt, dem sie ihre eigenen Erfahrungen gegenüberstellen können.

1.2.2. Vorgehen und Methodik

Abbildung 2 gibt einen Überblick über das Vorgehen und die Methodik bei der Erstellung der Studie.

Recherche und Wissensaufbereitung bilden den Start und die Grundlage. Es wurden bestehende **wissenschaftliche Publikationen** und das Anwendungswissen des Fraunhofer IAO, das im Rahmen von RAG-Projekten gesammelt wurde, genutzt, um die Funktionsweise von RAG darzustellen. Relevante Komponenten wurden erläutert und unterschiedliche Ansätze verglichen.

RAG-Prototypen, die durch das KI-Fortschrittszentrum gefördert wurden, werden als konkrete Beispiele für Anwendungsfälle aufbereitet und ausgewertet.

Weitere Daten zu Projekten mit umgesetzten RAG-Produktivsystemen wurden mittels **leitfadengestützter Interviews** erhoben. Bei der Methode des Leitfadeninterviews (Loosen, 2016) werden vorab Themen und Fragen festgelegt, die als Struktur dienen und dafür sorgen, dass die Interview-Ergebnisse vergleichbar bleiben. Eine flexible Reaktion auf die Gesprächssituation bleibt jederzeit möglich. Die offenen Fragen müssen nicht alle gestellt oder beantwortet

Abbildung 2: Überblick über Vorgehen, Methodik und Aufbau der Studie.

werden. Der hier genutzte Interview-Leitfaden befindet sich im Anhang der Studie. Am Schluss enthält er vier Thesen, zu denen die Zustimmung (nein, teilweise, ja) der interviewten Personen abgefragt wurde. Die Interviews fanden per Videokonferenz statt und dauerten zwischen einer und anderthalb Stunden. Der Interview-Leitfaden lag den interviewten Personen vorab vor, eine Vorbereitung wurde jedoch nicht erwartet. Alle Personen waren an der Umsetzung der RAG-Projekte in ihrer Organisation direkt beteiligt. Während des Interviews wurden wichtige Aussagen notiert. Die Interview-Aufzeichnungen und KI-Transkripte wurden genutzt, um diese Notizen im Nachgang zu ergänzen. Das auf diese Art und Weise gewonnene Praxiswissen wurde anonymisiert aufbereitet, ausgewertet und in Kapitel 4 dokumentiert.

Zur Ableitung der Handlungsempfehlungen haben die Autorin und der Autor drei weitere RAG-Experten des Fraunhofer IAO im Rahmen von zwei **Workshops** hinzugezogen. Die im Anschluss formulierten Handlungsempfehlungen gingen durch ein **weiteres Review aller Beteiligten.**

1.2.3. Untersuchte Anwendungsfälle

Im KI-Fortschrittszentrum wurden durch Fraunhofer IAO mit dem RAG-Ansatz im Rahmen von Quick Checks und einem Exploring Project mehrere **Prototypen** erstellt, die konkretes Erfahrungswissen generiert haben und zur Veranschaulichung in Kapitel 3 dargestellt werden. Für das Unternehmen Der Trendbeobachter ging es um die Erkennung und Bewertung von Zukunftsthesen. Mit Dr. Fritz Faulhaber GmbH & Co KG wurde ein RAG-System zur Unterstützung der Mitarbeitenden bei der Bearbeitung und Beantwortung von Kundenanfragen erprobt. Die Wissensextraktion aus Patentdokumenten war ein RAG-Anwendungsfall für Rackette Patentanwälte Part G mbB. Bei Stahl Cranesystems ging es um die Unterstützung bei der Auftragsabwicklung. Darüber hinaus bringt das Fraunhofer IAO Erfahrung mit einem RAG-Piloten für das Prüfungsamt der Pädagogischen Hochschule Ludwigsburg ein.

Mittels leitfadengestützter Interviews wurden Projekte zu **RAG-Produktivsystemen** in fünf größeren Organisationen untersucht. Es handelt sich dabei um Anwendungsfälle in einer Forschungsorganisation, bei zwei Versicherungsunternehmen, einer Krankenkasse und einem Automobilkonzern.

1.2.4. Gegenstand der Untersuchung

Die Studie betrachtet die praktische Umsetzung von RAG in Unternehmen. Die Untersuchungsdimensionen, die auch der Interview-Leitfaden berücksichtigt, sind die Anwendungsfälle und Zielsetzungen, die Projektorganisation und das Team, die Datengrundlage sowie die individuellen Erkenntnisse aus dem jeweiligen Projekt. Die genaue technische Umsetzung sowie Evaluierung der RAG-Anwendungsfälle wurden zwar erhoben, waren jedoch nicht Fokus der Untersuchung. Eine genaue Kategorisierung von RAG-Anwendungsfällen und eine quantitative Erhebung dazu, welche Anwendungsfälle in welchen Branchen wie oft umgesetzt werden, ist wünschenswert und möglicher Schwerpunkt zukünftiger Untersuchungen.

1.3. Aufbau der Studie

Kapitel 2 gibt einen Überblick über RAG als Architektur und Prozessprinzip zur wissensgestützten Textgenerierung durch Kopplung von zwei Technologien – einer Retrieval-Komponente und einem großen Sprachmodell (LLM). Beispiele von RAG-Anwendungsfällen im KI-Fortschrittszentrum und darüber hinaus werden in Kapitel 3 dargestellt. Kapitel 4 dokumentiert die Aussagen zur praktischen Umsetzung von RAG-Produktivsystemen aus den Interviews mit verschiedenen Organisationen. Das Ergebnis der Studie sind die Handlungsempfehlungen in Kapitel 5. Kapitel 6 fasst die Studie in einem Fazit zusammen und gibt einen kurzen Ausblick auf zukünftige Themen. Am Schluss der Veröffentlichung finden sich neben dem Interview-Leitfaden das Literaturverzeichnis und zusätzliche Informationen zu Fraunhofer, der Autorin und dem Autor sowie dem KI-Fortschrittszentrum.

2. Retrieval-Augmented Generation (RAG)

Ende 2024 hat OpenAI die Funktion eingeführt, dass alle Nutzerinnen und Nutzer mit ChatGPT direkt auf das Internet zugreifen können. Dies ermöglicht die Berücksichtigung von aktuellen Informationen. Davor konnte die Anwendung nur Informationen liefern, mit denen das Sprachmodell trainiert wurde und somit zum Beispiel nicht oder nicht korrekt auf Fragen wie »Wer ist aktuell der Präsident der USA?« antworten. Fragen, die spezifische, z. B. unternehmensinterne, Informationen benötigen, können damit auch nicht beantwortet werden.

Ein aus dem Prompt Engineering bekannter Ansatz ist die Möglichkeit, in einem Eingabetext (Prompt) relevante Kontextinformationen mitzugeben, z. B. ein oder mehrere Texte mit relevanten Informationen, die bei der Antwortgenerierung berücksichtigt werden sollen. Bei RAG geht es darum, spezifische Information in einem Datenpool, der meist nicht öffentlich zugänglich ist, zu identifizieren und bei der Antwortgenerierung zu berücksichtigen. Der Anwender muss die spezifischen Informationen vorher nicht kennen und auswählen, die Existenz dieser Daten ist ihm möglicherweise gar nicht bewusst. Ob die verwendeten Quellen bei der Antwort eines RAG-Systems angezeigt werden, kann für das System eingestellt werden.

Viele Mitarbeitende von Unternehmen haben Aufgaben, die internes Wissen erfordern. Sofern dieses Wissen implizit oder explizit in Daten vorhanden ist, kann ein spezifischer interner Chatbot ihre Arbeit in Zukunft unterstützen. Aktuell kommt bei der Umsetzung solcher spezifischer interner Chatbots häufig ein RAG-Ansatz zum Einsatz, der in diesem Abschnitt deswegen im Detail erklärt wird.

2.1. Einführung

Retrieval-Augmented Generation ermöglicht es, zielgerichtet das Kontextwissen vom Sprachmodellen (LLM) mit den pro Anfrage relevanten Informationen (Chunks) aus zuvor erstellten externen Wissensdatenbanken (meist Vektordatenbanken) anzureichern und so dem Sprachmodell die Informationen zur Beantwortung einer Anfrage direkt mitzugeben.

Der RAG-Ansatz wurde unter diesem Namen erstmals 2020 von (Lewis et al., 2020) für wissensintensive Aufgaben der natürlichen Sprachverarbeitung (NLP) beschrieben. RAG stärkt die sachliche Verankerung von LLMs und ergänzt sie um zusätzliches Wissen, ohne dass ein erneutes Modelltraining erforderlich ist. Der Ansatz reduziert in den meisten Fällen das Risiko von Halluzinationen und versucht gleichzeitig, sowohl die Ausgabequalität hoch als auch die Inferenzkosten[4] gering zu halten.

4 Kosten, die für den Betrieb eines KI-Modells bei einer Anfrage anfallen.

Die Ausgabe bezieht durch RAG aktuellen und ggf. domänenspezifischen Kontext mit ein, der nicht im Trainingssatz der LLMs enthalten ist. Zu beachten ist, dass auch RAG falsche Ergebnisse (durch Halluzinationen) nicht gänzlich verhindern kann.

Weiterhin basiert das »Wissen« der großen Sprachmodelle hauptsächlich auf öffentlich verfügbaren Daten. Dies macht sie im Allgemeinen zu guten »Allzweck-Antwortmaschinen«. Es fehlen ihnen jedoch so gut wie immer die spezifischen Informationen, die im Kontext des produktiven Einsatzes in einem Unternehmen tatsächliche Mehrwerte und Abgrenzungen gegenüber Standard-Antworten liefern. Es ist davon auszugehen, dass weitere große Qualitätssprünge der rein auf öffentlich zugänglichen Daten trainierten Modelle in Zukunft neue Ansätze benötigen.

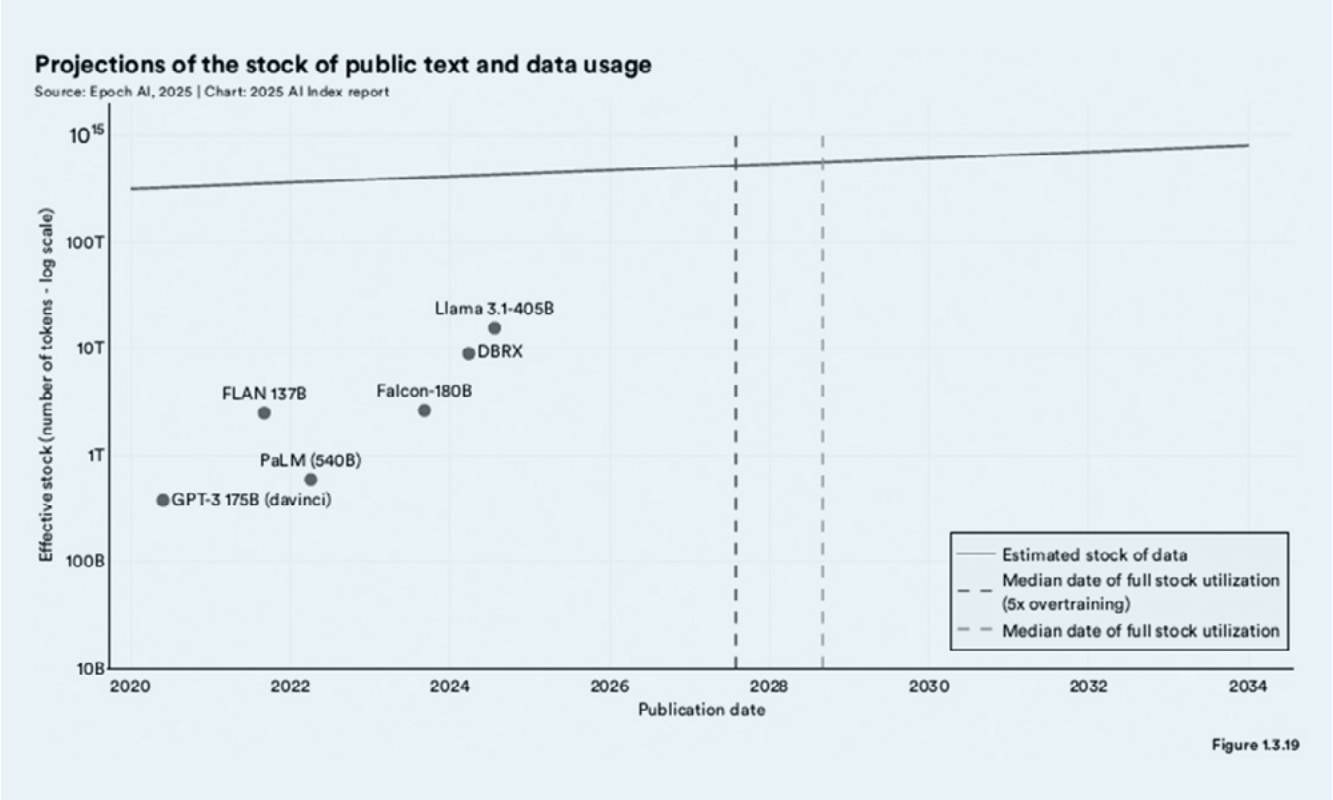


Abbildung 3: Abschätzung, wann die Menge öffentlicher Trainingsdaten aufgebraucht sein könnte (entnommen aus (Maslej et al., 2025) Figure 1.3.19)).

Abbildung 3 zeigt die Prognose von Epoch AI, nach der die Gesamtmenge an öffentlich verfügbaren Trainingsdaten im Web zwischen 2026 und 2032 erschöpft sein wird (Maslej et al., 2025). Eine weitere **Steigerung des Nutzens** ist möglich, indem Modelle **interne Daten** hinzuziehen. Ein technisch zugänglicher Ansatz dafür sind RAG-Systeme.

Mit dem Verweis auf die verwendeten Quelldokumente ermöglicht RAG – ähnlich zu Ansätzen zur Erklärbaren KI im klassischen Machine Learning (Schaaf, Wiedenroth & Wagner, 2021) – die Validierung der Ergebnisse durch die Nutzenden, was das **Vertrauen und die Akzeptanz in KI-Systeme** insgesamt erhöhen kann.

2.2. RAG-Einordnung und -Alternativen

Um Sprachmodelle an spezifische Kontexte und spezialisiertes Wissen anzupassen, gibt es alternative beziehungsweise parallele Ansätze: vom gesteuerten Anpassen der Eingabe (*Prompt Engineering*) über das Um- oder Nach-Trainieren des LLM selbst (*Fine-Tuning*), die Verwendung von Sprachmodellen mit langem Kontextfenster (*Long-Context*) bis zur Anbindung weiterer Systeme (*RAG*). Die Ansätze können sich teilweise stark in Komplexität und Auswirkungen auf das Systemverhalten unterscheiden.

Prompt Engineering ist der schnellste, kostengünstigste Hebel, um Antworten eines LLM anzupassen. Das zugrunde liegende Modell bleibt unverändert, lediglich die Art und Weise der Anfrage wird optimiert. Über gut gestaltete Prompts können bessere Instruktionen, Rollen, Beispiele, Formatvorgaben, usw. gesteuert werden. Der Ansatz ist besonders geeignet, wenn das Modell die Fakten schon »kennt« (also das benötigte Wissen im vortrainierten Modell grundsätzlich vorhanden ist) und es vor allem um Stil, Struktur, Tonalität und leichte Domänenanpassung geht. Benötigtes Faktenwissen für die Anfrage muss dem LLM gegebenenfalls als Input mit der Anfrage mitgegeben werden. Prompt Engineering erfordert kaum technische Implementierung, dafür aber ein systematisches Vorgehen und Evaluierung der Prompts. Des Weiteren können ausgeklügelte Prompts zum Teil sehr groß werden und zusammen mit dem benötigten Wissen einen großen Teil des Kontextfensters ausfüllen, was die Verarbeitungseffizienz der Modelle stark reduziert und die Kosten für die Verarbeitung drastisch erhöhen kann. Die Grenzen des Prompt Engineering sind dann erreicht, wenn das Kontextfenster nicht mehr ausreicht, um alle benötigten Informationen für die Anfrage aufzunehmen.

Fine-Tuning liegt am anderen Ende des Anwendungs- und Aufwands-Spektrums. Hierbei werden die Modellparameter selbst, typischerweise auf Basis von unternehmensspezifischen Beispieldaten, angepasst. Dies eignet sich besonders für stabile, klar umrissene Aufgaben – etwa wiederkehrende Textklassifikation, einen konsistenten Markenstil oder spezialisiertes Expertenvokabular – oder wenn Prompt Engineering allein nicht mehr ausreicht. Außerdem kann Fine-Tuning einen signifikanten Effizienzvorteil bringen. Das geforderte LLM-Verhalten oder der Kontext müssen nicht mehr bei jeder neuen Anfrage im Input mitgeschickt werden und die zu verarbeitende Token[5]-Anzahl sinkt. Ein Nachteil ist der signifikant höhere Aufwand bei Datenaufbereitung, Training sowie Monitoring. Außerdem ist eingelerntes Faktenwissen weniger flexibel aktualisierbar und kann veralten oder später sogar fehlerhaft sein.

RAG verfolgt das Ziel, einen Mittelweg zwischen der Komplexität und Starrheit von Fine-Tuning und der Beschränktheit und Ineffizienz von Prompt Engineering zu liefern. Neues oder dynamisches Wissen soll nicht »eingebrannt« werden, sondern aus Dokumenten, Datenbanken oder APIs abgerufen und dem Modell als Kontext bereitgestellt werden. RAG bietet ein gewisses Maß an Nachvollziehbarkeit, da die Wissensquellen, auf denen die Antwort basiert, zitiert werden können. Der Preis hierfür sind eine komplexere Systemarchitektur (Korpus, Retrieval, Qualitätssicherung) und zusätzliche Betriebsaufwände, die in den folgenden Abschnitten detaillierter beleuchtet werden.

Eine konzeptionelle Verortung der beschriebenen Ansätze ist in Abbildung 4 dargestellt.

Auch vor RAG gab es viele Ansätze zur Informationsbeschaffung (Information Retrieval (IR)) - eine Übersicht über alternative Ansätze liefert beispielsweise (Faloutsos & Oard, 1995).

5 LLMs verarbeiten und generieren Text nicht buchstaben- oder wortweise, sondern wandeln Wörter oder Wortteile in Zahlen um. *Token* sind eben jene kleinste Verarbeitungseinheit, in die LLMs Texte zerlegen.

Eine neuere Entwicklung, die mit der Skalierung moderner großer Sprachmodelle einher geht, ist die signifikante Vergrößerung des Kontextfensters großer Sprachmodelle in den Bereich der Millionen Token. Beispielsweise kann Gemini 3 Pro von Google (DeepMind, 2025) bis zu einer Million Input Token verarbeiten, was einer durchschnittlichen Menge von ca. 750 000 Wörtern entspricht. Dies ermöglicht, große Mengen Kontext direkt über den Prompt bei der Anfrage mitzuliefern, also die Grenzen von Prompt Engineering zu erweitern. Um mit solchen – tendenziell sehr teuren – Anfragen effizienter umzugehen, werden Zwischenspeicherungsverfahren (Caching) angewendet, um bereits berechnete ähnliche Ergebnisse wiederzuverwenden. Diese **Long-Context- und Cache-Augmented-Ansätze** versuchen darüber, die RAG-Logik teilweise zu ersetzen, haben aber wiederum eigene Mängel wie z. B. den graduellen Leistungsabfall der Genauigkeit von Modellen beim Ausnutzen größerer Teile des Kontextfensters (Fraga, 2024) (Hong, Troynikov, & Huber, 2025).

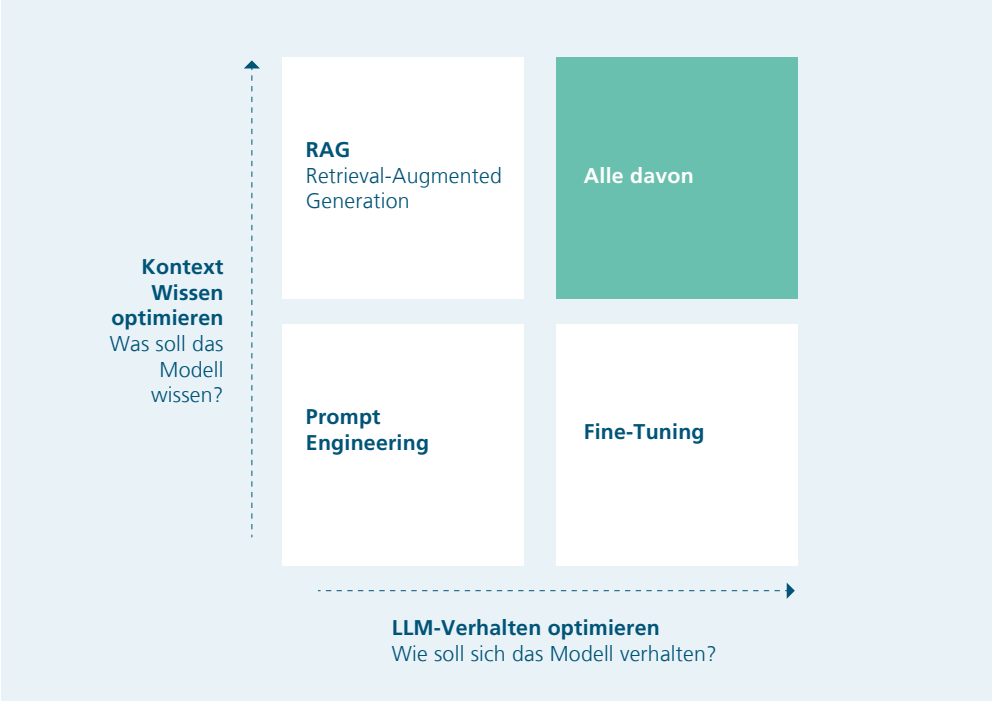


Abbildung 4: Einordnung der verschiedenen Ansätze, um Sprachmodelle an gefordertes Verhalten und Kontext anzupassen. Die Ansätze unterscheiden sich in Komplexität und Auswirkung auf das LLM-Verhalten sowie in der Leistungsfähigkeit. Grafik in Anlehnung an (OpenAI, 2023).

2.3. Naive RAG

Bei der technisch ursprünglichen Umsetzung von RAG (hier *Naive RAG* genannt) wird dem Sprachmodell zusätzliches Wissen über den Kontext mit einer minimalen Menge weiterer Komponenten zur Laufzeit übergeben (Lewis et al., 2020).

Abbildung 5 zeigt die grundsätzliche Funktionsweise von RAG[6]: Als Grundlage erfolgt der Aufbau eines Korpus. Eine spezielle Art von Wissensdatenbank mit relevanten Inhalten wird erstellt, in der Informationen über ein sogenanntes Embedding-Modell meist als hochdimensionale numerische Werte (Vektoren) abgebildet werden. Im Betrieb wird im Korpus nach relevanten Inhalten zu gestellten Anfragen gesucht und das Ergebnis wird dem Sprachmodell zur Antwortgenerierung in Form eines Prompt-Templates zusammen mit der Frage übergeben.

6 Der Einfachheit halber klammert der dargestellte Workflow die weiterführenden Techniken und SOTA-Erweiterungen zugunsten einer klareren Darstellung aus.

Der Ablauf der Ausführung bei Naive RAG lässt sich damit in folgende Schritte unterteilen (Oche, Folashade, Ghosal, & Biswas, 2025):

1. **Anfragen-Kodierung**: Der Anfrage-Text wird mit dem gleichen Ansatz in dieselbe numerische Repräsentation umgewandelt wie auch die Dokumente beim Korpusaufbau. Generell gibt es mehrere Kodierungs-Möglichkeiten[7]. Meist wird ein vorab trainiertes Embedding-Modell genutzt, das möglichst viele semantische Informationen des übergebenen Texts erfasst.

2. **Dokumenten-Abruf**: Im Retrieval-Schritt wird mit der numerischen Repräsentation der Anfrage eine Suche auf der Datenbank durchgeführt. Ein vorab definiertes Ähnlichkeitsmaß dient hierbei als Metrik, um die *K* relevantesten Dokumententeile (Chunks) zu identifizieren und zurückzuliefern.

3. **Kontext-Anreicherung**: Das Prompt Template wird sowohl mit den aus dem Retrieval-Schritt identifizierten Chunks als auch mit der ursprünglichen Anfrage befüllt.

4. **Antwort-Generierung**: Je nach genauer technischer Umsetzung generiert das eingebundene LLM entweder für jeden Chunk eine eigene Antwort und es obliegt dem RAG-System, die passende finale Antwort auszuwählen (*late fusion*), oder alle abgerufenen Dokumententeile werden zusammen verarbeitet und eine Antwort wird aus dem zusammenhängenden Text aller Chunks generiert (*early fusion*).

Naive RAG hat Grenzen in Genauigkeit und Flexibilität – es fehlen beispielsweise Mechanismen zur Qualitätskontrolle der Dokumente oder Rückmeldungen an den Retriever. Außerdem können komplexere Daten-Modalitäten abseits von reinem Text nicht verarbeitet werden. Um diese Einschränkungen zu adressieren, gab es im Laufe der letzten Jahre diverse Weiterentwicklungen und Verbesserungen des naiven Ansatzes.

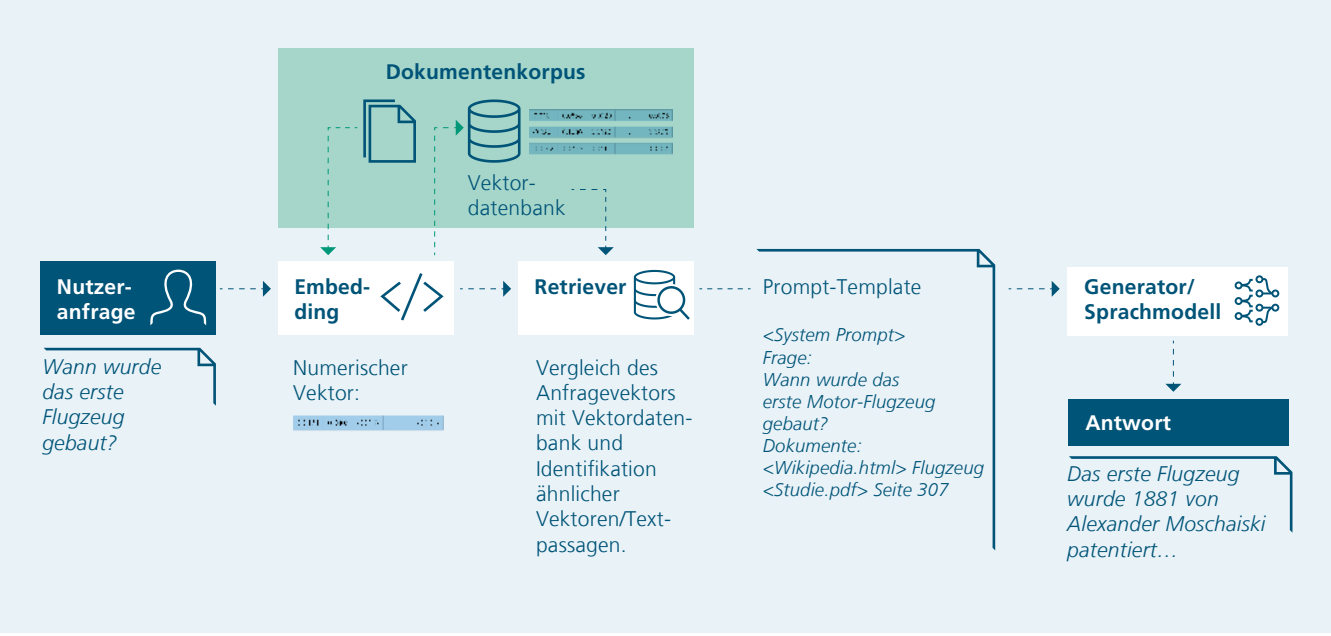


Abbildung 5: Workflow für Naive RAG. Zu Beginn wird eine Wissensdatenbank erstellt (Korpusaufbau). Im Betrieb wird nach ähnlichen Inhalten zu den gestellten Anfragen gesucht, die dann als Kontext für die Antwortgenerierung verwendet werden. Quelle: eigene Darstellung.

7 Es gibt »*sparse*« und »*dense*« Embeddings sowie hybride Ansätze, die sich hauptsächlich in der Art des verwendeten Algorithmus unterscheiden. *Sparse Embeddings* nutzen z. B. Schlagwortsuche (BM25/TF-IDF) auf den reinen Texten, während bei *dense Embeddings* trainierte Modelle diese in hochdimensionale Vektoren überführen.

2.4. RAG-Arten und Weiterentwicklungen

Weiterentwicklungen und Verbesserungen von RAG adressieren Schwächen im Ansatz oder bei der Interaktion. Sie bieten alternative Techniken bei der Einbindung der Daten oder haben Autonomie im Fokus.

2.4.1. Erweiterte RAG

Bei der Verwendung von Naive RAG (siehe Abschnitt 2.3), welches meist eine einfache Vektorsuche mit dem Nutzer-Input als Ausgangspunkt durchführt, gibt es mehrere potenzielle Schwachstellen, die die Genauigkeit der Systeme beeinflussen: Unscharf formulierte oder zu knappe Benutzeranfragen, mehrfach auftretende oder semantisch ähnliche Einträge im Korpus zu unterschiedlichen Entitäten/Themen, das Auffinden von relevanten und irrelevanten Chunks im selben Retrieval-Schritt usw. Um diese Schwächen zu adressieren, gibt es mehrere Techniken, die vor und nach dem Abfrage-Schritt angewendet werden, um sowohl den Suchumfang als auch die Relevanz der Ergebnisse zu optimieren.

Bei **Query Expansion** wird die ursprüngliche Nutzeranfrage automatisch um zusätzliche Begriffe erweitert (Synonyme, verwandte Konzepte, häufig mitrecherchierte Terme etc.), um mehr relevante Dokumente zu finden und Wort-Mismatch zu verringern (Carpineto & Romano, 2012). **Query Rewriting** (Riezler & Liu, 2010) formuliert die Anfrage in gewisser Form um, sodass sie besser zur Art des Inhalts in der Wissensdatenbank passt. Das kann von simpler Übersetzung der Anfrage in die Zielsprache des Korpus bis hin zur Generierung von Pseudo-Antworten auf die ursprüngliche Frage reichen, um dem Embedding-Modell die Assoziation zwischen Anfrage und relevanten Inhalten im Korpus zu erleichtern (Ma, Gong, He, Zhao, & Duan, 2023). Beide Techniken finden *pre-retrival* statt, werden also direkt auf der Anfrage angewendet.
Im Gegensatz dazu ist **Reranking** ein zweistufiges *post-retrieval* Verfahren (vgl. (Dang, Bendersky, & Croft, 2013)), bei dem zuerst eine größere Anzahl an Chunks als Kandidatenliste im Dokumenten-Abruf-Schritt zurückgeliefert und anschließend mittels eines zweiten Modells nach Relevanz bewertet und sortiert werden (Kim & Lee, 2024).

Eine schematische Darstellung der Ansätze ist in Abbildung 6 dargestellt.

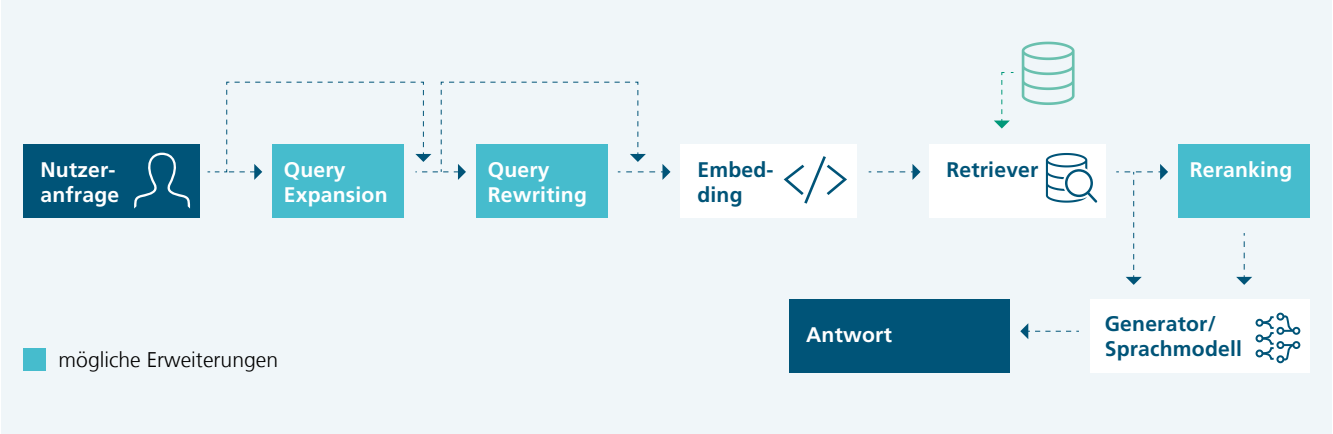


Abbildung 6: Konzeptioneller Ablauf einer erweiterten RAG Pipeline mit den zusätzlichen Techniken Query Expansion, Query Rewriting und Reranking. In Anlehnung an (Gao et al., 2023) Fig. 3.

2.4.2. Multimodale RAG

Multimodale RAG erweitert das klassische System, indem verschiedene Datenmodalitäten – nicht nur Text, sondern z. B. auch Bilder, Audio oder strukturierte Tabellen – in den Abruf- und Generierungsprozess einbezogen werden (Abootorabi et al., 2025). Dies ermöglicht, Anfragen zu beantworten, die über rein textbasierte Informationen hinausgehen. Ein multimodales RAG-System kann zum Beispiel zusätzlich zum reinen Text auf Bilddatenbanken, Audio-Transkripte oder Tabellen zugreifen.

Die Multimodalität kann sowohl für die Art des Inputs (wenn die Anfrage mehrere Datenarten enthält) als auch für die Suche auf dem Korpus (wenn die Wissensdatenbank verschiedene Dateiformate enthält) oder für beides realisiert werden[8].

Die Zusammenführung der verschiedenen Modalitäten für die Generierung kann auf mehrere Arten erfolgen: entweder durch eine vorgeschaltete Transformation aller Inhalte in Text oder durch ein multimodales Sprachmodell, das Bilder und andere Formate direkt verarbeiten kann. Multimodale RAG-Systeme sind deshalb deutlich anspruchsvoller in Aufbau und Betrieb. Sie müssen nicht nur verschiedene KI-Modelle orchestrieren, sondern auch sicherstellen, dass die Ergebnisse der unterschiedlichen Modalitäten sinnvoll zusammengeführt werden. Anwendungsgebiete liegen beispielsweise in Domänen wie visuellem Fragen-Antworten (QA), medizinischer Diagnose (Bild + Text) oder Multimedia-Suchmaschinen, in denen Text und Bild/Video kombiniert ausgewertet werden.

2.4.3. Graph-RAG

Graph-basierte Retrieval-Augmented Generation (Graph-RAG) ist eine weitere komplexe Ausprägung von RAG und integriert strukturierte Wissensgraphen (Daten- und Assoziations-Relationen) in den Abrufprozess (Hu et al., 2025) (Han et al., 2024). Dies hat den Vorteil, dass Beziehungen zwischen Informationen, die semantisch weit voneinander entfernt liegen, einfacher oder überhaupt erst im Retrieval-Schritt gemeinsam gefunden werden können. Das ist vor allem dann sinnvoll, wenn Transferleistungen oder Multi-Hop-Abrufe notwendig sind, bei denen mehrere Quellen – teilweise iterativ – zur Beantwortung einer Anfrage durchsucht werden müssen. Ein Beispiel hierfür ist die Frage nach einer bestimmten Produkteigenschaft oder -eignung, die nicht explizit im hinterlegten Produktdatenblatt beschrieben ist, sondern stattdessen auf eine zugehörige Zertifizierungsnorm verweist. Die Transferleistung, von der dokumentierten Norm im Produktblatt auf den Abruf der Norm-Beschreibung zu schließen, die erst einmal nichts mit dem Produkt an sich zu tun hat, können andere RAG-Systeme meist nicht erbringen.

Die technische Umsetzung von Graph-RAG ist komplex, da sowohl bei der Erstellung zuerst ein Wissensgraph gebaut und gepflegt als auch während des Betriebs ein geeigneter Sub-Graph passend zur Anfrage bestimmt und dem Generator (Sprachmodell) verständlich gemacht werden muss.

8 Wir klammern hier bewusst die Möglichkeit der multimodalen Output-Generierung aus, da wir die Fähigkeit, mehrere Datenarten zu erzeugen, dem generativen Modell zuschreiben – unabhängig von RAG.

2.4.4. Agentische RAG

Die meisten RAG-Arten bilden ein starres Vorgehen von einmaligem Abruf und Generierung pro Anfrage ab. Agentische RAG (Singh, Jamdar, & Kaul, 2025) nutzt KI-Agenten in der RAG-Pipeline als zentrales Steuerelement, um dynamischere und komplexere Wissensabruf-Strategien zu ermöglichen. Dazu steuert ein Agent (typischerweise ein LLM mit Planungs- und Tool-Fähigkeiten) den Prozess. Dieser Agent kann eigenständig Entscheidungen treffen (z. B. etwa welche Wissensquelle oder welches Tool als nächstes genutzt werden soll) und so mehrschrittige Workflows ausführen. Dadurch steigt die Anpassungsfähigkeit und Genauigkeit des Systems: Das LLM kann Informationen aus mehreren Quellen sukzessive einholen, Zwischenergebnisse speichern und auswerten sowie externe Werkzeuge (z. B. Rechner, Web-Suchen oder Datenbankabfragen) einbinden, bevor es die finale Antwort formuliert. Agentische RAG kann somit Aufgaben lösen, die mit einem einfachen einmaligen Abruf oder ohne weitere Transferleistungen nicht lösbar wären.

Agentische RAG-Systeme sind neuartige, sehr komplexe RAG-Varianten und derzeit Gegenstand aktiver Forschung und Weiterentwicklung. Wie die meisten Agenten-Systeme sind sie meist schwerer kontrollier- und absicherbar, dafür aber auch (je nach Ausprägung und verfügbaren Tools) generalisierbarer und mächtiger als andere Varianten.

2.5. Gegenüberstellung der RAG-Arten

Im Folgenden werden die vorgestellten RAG-Arten in Relation gesetzt und deren Vor- und Nachteile kurz dargestellt. Die Tabelle erhebt keinen Anspruch auf Vollständigkeit und soll als Möglichkeit für eine erste Einschätzung der verschiedenen RAG-Ansätze dienen:

Tabelle 1: Gegenüberstellung verschiedener RAG-Ausprägungen.

RAG-Art	Verwendete Datenart	Komplexität	Vorteile	Nachteile	Beispiel-Frameworks[9]
Naive RAG	Text	**Gering** – minimale Pipeline (Embedding-Modell + Vektordatenbank + einfaches Prompting)	Einfache Implementierung Einbettung verifizierbarer Kontextdokumente in den Prompt	Begrenzte Genauigkeit/ Flexibilität Keine Qualitätskontrolle Nur Textdaten Aufwendiger als naive RAG	LangChain LlamaIndex Pinecone* Weaviate* Haystack
Erweiterte RAG	Text und ggf. Metadaten	**Mittel** – zusätzliche Schritte zur Abfrage-Optimierung vor/nach dem Retrieval	Näher an State-of-the-Art-Systemen Höhere Trefferqualität Filtern irrelevanter Chunks	Risiko von schlechten Query Expansions Zusätzliche Modelle (z.B. Reranker) nötig	LangChain LlamaIndex RAGFlow Cohere*
Multimodale RAG	Text, Bild, Audio, strukturierte Daten	**Hoch** – orchestriert unterschiedliche KI-Modelle je nach Modalität; aufwendige Cross-Modal-Fusion der Ergebnisse erforderlich	Kann Informationen aus mehreren Datentypen extrahieren Sehr flexibel einsetzbar	Komplexer Aufbau: benötigt mehrere spezialisierte Modelle und Abstimmung oder mächtige multimodale Modelle Höherer Rechenaufwand (Kosten)	Google Vertex AI* Microsoft Azure AI* OpenCLIP LangChain LlamaIndex
Graph-RAG	Text und Wissensgraph mit Entitäten/Kanten	**Sehr hoch** – benötigt Aufbau und Pflege eines Wissensgraphen und komplexe Pipeline	Erweitert die Leistungsfähigkeit anderer RAG-Ansätze Erhöhte Transferleistung bei der Beantwortung komplexer Fragen Beantwortung von Multi-Hop-Fragen möglich (kann semantisch weit entfernte Fakten verknüpfen)	Sehr aufwendige Korpus-Erstellung: Erzeugung und Pflege des Wissensgraphen ist aufwendig Komplexe Laufzeitlogik (Identifikation des Sub-Graphs, Integration in Prompt etc.) Technisch komplex	LangGraph Microsoft GraphRAG Neo4j* FalkorDB*
Agentische RAG	Beliebige Datenquellen dynamisch über Tools eingebunden (Web, DBs, Graphen, ...)	**Sehr hoch** – komplexer, mehrschrittiger Ablauf: autonomer LLM-Agent plant Queries und Aktionen, ruft iterativ Informationen ab und kombiniert Tools (Planungsschleifen, Zwischenanalysen, Speicher)	Hohe Flexibilität und Autonomie Kann planen und schlussfolgern Kann dynamisch nach Bedarf Datenquellen einbinden Kann auf Kontext-Wechsel und veränderte Bedingungen reagieren	Schwer kontrollier- und absicherbar Hohe Latenz und Kosten bei der Ausführung Technisch komplex Benötigt vorsichtiges Prompt-Design für zuverlässige Ergebnisse	LangGraph CrewAI Microsoft AutoGen OpenAI Agents SDK ElevenLabs* Google Agent Development Kit (ADK)* AWS Bedrock* Azure AI Foundry*

9 Viele Frameworks wie LangChain oder LlamaIndex bieten Unterstützung für mehrere RAG-Arten, sind hier aber an den Stellen gelistet, wo sie derzeit aus Sicht der Autoren ihre Stärken haben. Mit * markierte Frameworks/Tools sind proprietär (bezogen auf die Lizenz).

3. Anwendungsfälle für RAG

Der folgende Abschnitt gibt zunächst eine generelle Einordnung von möglichen Anwendungsfällen, bei denen der Einsatz von RAG einen Mehrwert liefern kann. Des Weiteren liefert das Kapitel einen Einblick über ausgewählte Pilotprojekte mit RAG-Bezug sowie deren Zielstellung, Umsetzung und Nutzen, die durch das Fraunhofer IAO in Zusammenarbeit mit Praxispartnern umgesetzt wurden. Dabei stehen ausgewählte freigegebene Anwendungsfälle aus dem KI-Fortschrittszentrum im Fokus. Einige dieser Proof of Concepts (PoCs) nutzen nur einen Teil der gesamten RAG-Pipeline, was zeigt, dass bereits Teilaspekte einen unternehmerischen Mehrwert für bestimmte Prozesse liefern können.

3.1. Typen von Anwendungsfällen

Im Folgenden werden exemplarisch einige beispielhafte RAG-Anwendungsfälle aus unterschiedlichen Bereichen vorgestellt. Die Einordnung soll dabei helfen, eigene RAG-Ideen zu verorten und einzuschätzen. Die Bewertung der Komplexität ergibt sich aus der Beurteilung der Fraunhofer-Experten zum geschätzten Implementierungsaufwand und Annahmen der benötigten Schnittstellen/Integrationen zu anderen Systemen. Eine genauere Beurteilung ist nur für konkrete Einzelfälle und Rahmenbedingungen möglich.

3.1.1. Interne Effizienz und Wissensmanagement

Diese Anwendungsfälle sind meist intern, womit die Auswirkungen von Fehlern (z. B. Halluzination) häufig geringer sind. Außerdem liegt oftmals ein Großteil der benötigten Datenquellen bereits im Unternehmen vor, wodurch sich die Use Cases meist gut für erste Pilotprojekte eignen.

Tabelle 2: RAG-Anwendungsfälle für das interne Wissensmanagement.

Use Case	Kurzbeschreibung	Typische Quellen	Geschäftlicher Nutzen	Komplexität	Kritische Herausforderungen
Unternehmenseigener Wissens-Chatbot	Fragen zu internen Abläufen und dokumentiertem Wissen (»Google für die eigene Firma«)	Intranet (SharePoint/ Confluence), Wikis, PDFs…	Zeitersparnis bei der Recherche; Wissenssilos aufbrechen	Niedrig	Datenhaltung und -pflege: Veraltete Daten, »Garbage in, Garbage out«
Onboarding / HR-Assistenz	Fragen zu Benefits, Urlaub, Spesen, IT-Setup	HR-Richtlinien, Handbücher, Formulare	Entlastung von HR; Schnelleres »Ramp-up« neuer Mitarbeitender	Mittel	Einführung und Einhaltung von Zugriffsrechten (ACLs)
Meeting- / Wissensassistent	Protokolle und Kontext abfragen	Transkripte (Teams/ Zoom), Notizen, Jira-Tickets	Nachvollziehbarkeit und Dokumentation von Entscheidungen	Hoch	Intelligentes Chunking und Zusammenfassungen; komplexere Datenvorverarbeitung und Zuordnung

3.1.2. Kundenorientierung und Support

Diese RAG-Systeme haben meist einen höheren ROI, allerdings auch hohe Anforderungen an Qualitätssicherung und Zuverlässigkeit. Die Lösung zeigt sich nach außen gegenüber Kunden, woraus sich Reputationsrisiken ergeben.

Tabelle 3: RAG-Anwendungsfälle, die nach außen gerichtet sind.

Use Case	Kurzbeschreibung	Typische Quellen	Geschäftlicher Nutzen	Komplexität	Kritische Herausforderung
Kundenservice-Assistenz	KI schlägt Support-Mitarbeitenden Antworten inkl. Quellen vor (Human-in-the-loop)	Wissensdatenbank, Ticket-Historie, FAQs	Schnellere Ticketlösung; konsistentere Antworten	Mittel	Aktualität: Support-Wissen ändert sich täglich
Technischer Support/Field Service	Unterstützung bei komplexen Fehlern, Fehlercodes und Reparaturen	Maschinen-Handbücher, Schaltpläne, Logs, Betriebsanleitungen	Schnellere Problemlösung vor Ort; Expertenwissen für Junioren verfügbar	Hoch	Multimodalität: Oft müssen Diagramme oder Screenshots mit interpretiert werden
Produkt-/ API-Doku-Chat	Entwickler fragen nach Produkt/ Funktion und erhalten Code	API-Docs (Swagger/ OpenAPI), Markdown-Readmes, GitHub	Bessere Developer Experience (DX); Entlastung des Dev-Supports.	Hoch	Code-Kontext und -Generierung; Mehrere valide versionierte Schnittstellen

3.1.3. Spezialisierte Fachbereiche

Spezialisierte RAG-Lösungen für bestimmte Unternehmensbereiche bieten oft den größten individuellen Mehrwert, sind aber meist nur schwer generalisier- oder übertragbar. Häufig sind RAG-Ansätze hier ein Bestandteil größerer Systeme. Teilweise sind Modellanpassungen an die Domäne oder strikte Compliance-Regeln nötig.

Tabelle 4: RAG-Anwendungsfälle, die domänenspezifisch sind.

Use Case	Kurzbeschreibung	Typische Quellen	Geschäftlicher Nutzen	Komplexität	Kritische Herausforderung
Compliance & Legal	Prüfung von Verträgen oder Fragen zur Regulatorik (mit exakten Zitatnachweisen)	Gesetzestexte, Policies, Verträge, Verordnungen	Risikominimierung; Beschleunigung der Vertrags- und Dokumentenerstellung	Hoch	Exaktheit bzw. Minimierung von Halluzinationen; Einhaltung sprachlicher Compliance-Regeln
Vertriebs-/ Angebots-Bot	Automatisches Ausfüllen von Ausschreibungen oder Erstellung von Angeboten	Frühere Angebote, erfolgreiche Ausschreibungen, Produktdaten	Skalierung des Vertriebs; schnelleres Reagieren auf Ausschreibungen	Mittel	Schutz interner Daten und IP; Einhaltung vorgegebener Strukturen und Templates
Data-/ BI-Assistent	Kontextualisierung, Verknüpfung und Analyse strukturierter Daten	Datenkataloge, Glossare, BI-Dashboards (Metadaten), Datenbanken	Einfachere und schnellere Insight-Generierung; Entlastung des Data-Teams	Hoch	Strukturierte vs. unstrukturierte Daten; oft komplexes Verbund-System

3.2. Quick Check: Erkennung und Bewertung von Zukunftsthesen

Im Rahmen des Quick Checks[10] am KI-Fortschrittszentrum »Erkennung und Bewertung von Zukunftsthesen« wurde für das Unternehmen Der Trendbeobachter untersucht, inwiefern KI-gestützte Analysen großer Textmengen die Arbeit bei der Identifikation neuer Trends und der Validierung bestehender Zukunftsthesen unterstützen können.

Bisher wurden Trendanalysen überwiegend durch manuelle Recherche in allgemeinen und fachspezifischen Quellen erstellt. Dieses Vorgehen ist qualitativ hochwertig, jedoch zeitintensiv und bei stetig wachsenden Datenmengen nur sehr begrenzt skalierbar. Der Quick Check adressierte die Frage, wie sich eine bereits existierende, große Datenbasis systematisch auswerten lässt, ohne die inhaltliche Einordnung und Beratungskompetenz des Unternehmens zu ersetzen.

Als Lösungsidee wurden zwei KI-Assistenzfunktionen prototypisch entwickelt:

1. Erschließung neuer, bislang unbekannter Trends aus einer großen Anzahl aktueller Quellen (z. B. Studien, Zeitungs- und Fachartikel).

2. Prüfung bzw. Validierung bereits formulierter Thesen auf dieser Datenbasis.

Methodisch kam nur der erste Teil einer vollständigen RAG-Pipeline zum Einsatz: Es wurde der Retrieval-Teil genutzt, um aus dem umfangreichen Textkorpus diejenigen Passagen zu identifizieren, die für ein Thema oder eine These besonders relevant erscheinen. Die generative Texterstellung durch LLMs wurde nicht benötigt. Darauf aufbauend erfolgte die Analyse über zwei unterschiedliche, nicht-generative Verfahren: Für die Trendfindung wurde unüberwachtes Clustering angewendet, um thematische Häufungen und Veränderungscluster zu erkennen. Für die Thesenvalidierung wurde eine semantische Ähnlichkeitsanalyse genutzt, um Textstellen zu finden, die eine These stützen, präzisieren oder ihr widersprechen. Schematisch ist dies in Abbildung 7 dargestellt.

Zielsetzung des Anwendungsfalls ist, eine deutlich höhere Geschwindigkeit und Reichweite bei der Trendbeobachtung zu erreichen. Durch die automatisierte Auswertung kann das Unternehmen früher entstehende Trends erkennen und darauf reagieren. Gleichzeitig erhöht die systematische Suche die Chance, Themen aufzudecken, die in manueller Recherche übersehen werden.

Mehr Informationen zum Projekt:

www.ki-fortschritts-zentrum.de/projekt/erkennung-und-bewertung-von-zukunftsthesen

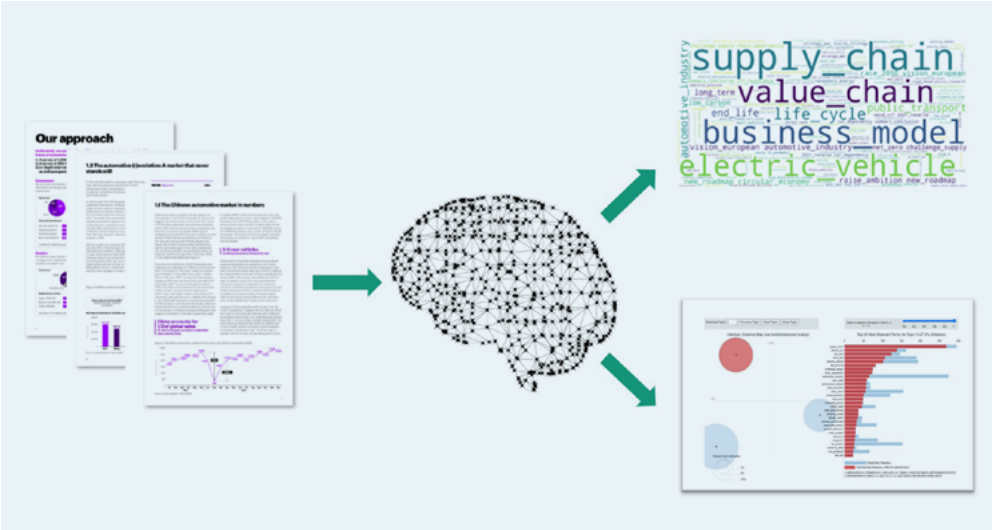


Abbildung 7: Darstellung der Erkennung und Validierung von Zukunftstrends aus Medienberichten für den PoC »Erkennung und Bewertung von Zukunftsthesen« mit der Firma Der Trendbeobachter. Quelle: eigene Darstellung.

10 Ein Quick Check ist eines von mehreren Formaten des KI-Fortschrittzentrums, in dem Innovationsideen analysiert und gemeinsam Lösungsansätze erarbeitet werden. Siehe auch: https://www.ki-fortschrittszentrum.de/zusammenarbeit

3.3. Quick Check und Exploring Project: AI-Powered Assistance for Industrial Drive Systems

Zusammen mit der Firma Dr. Fritz Faulhaber GmbH & Co KG wurde in einem Quick Check und fortführenden Exploring Project ein Assistenzsystem für die Support- Mitarbeitenden zur Bearbeitung und Beantwortung von Kundenanfragen erstellt. Konkret wurden Anwenderdokumentationen von komplexen Antriebssystemen erschlossen.

Die Support-Anfragen sind in der Regel kontextabhängig und entwickeln sich oft über mehrere E-Mails hinweg. Für eine korrekte Bearbeitung müssen die Mitarbeitenden relevante Informationen (z. B. Controller-Typ, Seriennummer, Softwareversion, Motor) aus der Kommunikation erfassen und mit Produktwissen verknüpfen. Die dafür genutzten Informationsquellen sind beispielsweise Handbücher, Funktionstabellen, AppNotes sowie interne Hilfsmitteln wie Kurznotizen, Checklisten, Textbausteine, Fragebäume usw. Die Verteilung dieses Wissens auf viele Dokumente erhöht die Komplexität und kann die Reaktionszeiten verlängern.

Das entwickelte Pilotsystem unterstützt bei der automatischen Kontextklärung und ermöglicht darauf aufbauend die Beantwortung einfacher Anfragen. Dazu extrahiert die Lösung zentrale Informationen aus E-Mails, ordnet die Anfrage über Mapping-Tabellen einer Produktfamilie zu und klassifiziert sie entlang typischer Fragegebiete bzw. Phasen im Produktlebenszyklus (z. B. Inbetriebnahme, Integration, Programmierung, Errorhandling). Daraus wird eine Zusammenfassung erstellt und zum schnellen Wiederauffinden verschlagwortet. Außerdem werden bestimmte fehlende Informationen erkannt und damit Vorschläge für gezielte Rückfragen erstellt. Darauf aufbauend wird die Wissensbasis genutzt, um relevante Dokumente vorzuschlagen, und über einen RAG-Chatbot werden passende Antworten zur Anfrage generiert.

Der PoC wurde mit einem KI-Workflow-Tool (siehe Abbildung 8) implementiert und dient dazu, die Machbarkeit, Qualität und Entlastungseffekte zu bewerten. In der Erprobung werden verkürzte Bearbeitungszeiten, konsistentere Kontextklärung und eine spürbare Verbesserung der Antwortgeschwindigkeit und -präzision für industrielle Kunden erwartet.

Mehr Informationen zum Projekt:

https://www.ki-fortschrittszentrum.de/projekte/

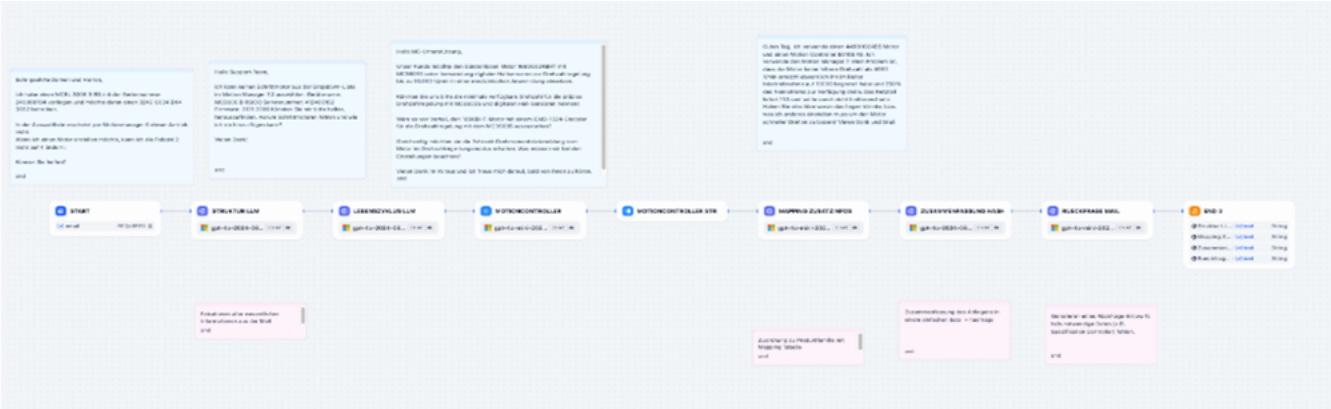


Abbildung 8: Exemplarische Pipeline Implementation des PoC in einem graphischen KI-Workflow-Tool. Quelle: eigene Darstellung.

3.4. Quick Check: AI-Patentanmeldung – Wissensextraktion aus Patentdokumenten mit KI

Mehr Informationen zum Projekt:

www.ki-fortschritts-zentrum.de/projekt/ai-patentanmeldung

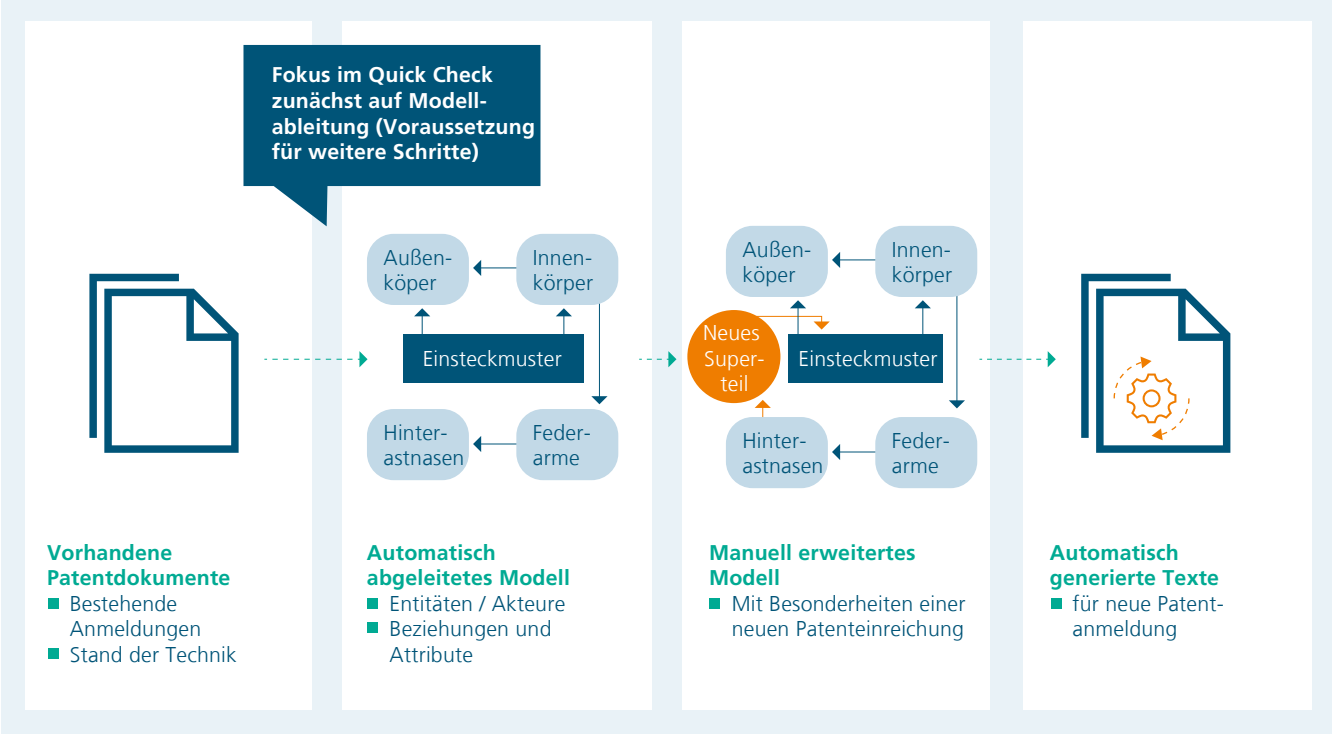
Im Rahmen des Quick Checks »AI-Patentanmeldung« für Rackette Patentanwälte PartG mbB wurde im KI-Fortschrittszentrum untersucht, wie sich Inhalte aus Patentdokumenten mithilfe von KI so aufbereiten lassen, dass sie die Erstellung neuer Anmeldungstexte künftig unterstützen.

Ausgangspunkt war, dass Patentanmeldungen heute überwiegend manuell durch hochspezialisierte Fachkräfte erstellt werden. Dabei wird ein neuer Text typischerweise auf Basis vorhandener Patentschriften formuliert, was umfangreiche Recherche und hohe Aufmerksamkeit für formale Anforderungen erfordert. Rackette bringt hierfür langjährige Erfahrung und fundiertes Domänenwissen aus der Praxis ein.

Der Lösungsansatz soll das in Patentdokumenten enthaltene Wissen zunächst automatisiert extrahieren und in strukturierter Form bereitstellen. Dazu werden regelbasierte Methoden und unüberwachte KI-Ansätze als kombinierte Sprachverarbeitungs-Verfahren (NLP) eingesetzt. Ziel war nicht die direkte Texterstellung, sondern die Generierung eines Wissensgraphen: Entitäten (z. B. Komponenten, Verfahren, Materialien) und deren Beziehungen (z. B. »besteht aus«, »verbindet«, »wird verwendet für«) sollten aus typischen Formulierungen in Patenttexten abgeleitet werden. Im Prototyp wurden dazu Muster genutzt, um wiederkehrende Satzkonstruktionen zu erkennen und daraus Knoten und Kanten im Graphen aufzubauen. Ergänzend kamen Matching-Algorithmen zum Einsatz, die Beziehungen anhand des Satzkontexts über Ähnlichkeits-Metriken identifizieren. Darüber wurden Retrieval-Funktionen zur Wissensextraktion implementiert, die einen strukturierten Überblick über wesentliche Informationen und Beziehungen eines Dokuments liefern.

Der Fokus des PoCs sowie mögliche Weiterentwicklungen sind in Abbildung 9 dargestellt.

Abbildung 9: Zielvision im Quick Check für automatisierte Patentgenerierung. Quelle: eigene Darstellung.



3.5. Quick Check: KI unterstützte Auftragsabwicklung

Im Quick Check »KI unterstützte Auftragsabwicklung« mit dem Unternehmen Stahl Cranesystems im KI-Fortschrittszentrum wurde die Vereinfachung der Variantenkonfiguration von Hebezeugen untersucht. Der Fokus lag dabei auf der zuverlässigen Suche im existierenden Auftragsbestand, um damit neue Kundenanfragen schneller bearbeiten zu können. Kernannahme des Ansatzes war, dass für viele neue Aufträge bereits vergleichbare Lösungen im historischen Datenbestand existieren und als Referenz wiederverwendet werden können.

Das im Einsatz befindliche Variantenkonfigurationssystem überführt Kundenaufträge in Materialnummern und Auftragsstücklisten. Nicht im System abbildbare Merkmale werden häufig als Freitext erfasst. Diese Freitexte sind inhaltlich relevant, jedoch unterschiedlich formuliert. In der Praxis erschwert dies sowohl die zuverlässige Interpretation als auch die Suche nach passenden früheren Aufträgen, da klassische Textsuche stark von identischen Begriffen und Schreibweisen abhängt.

Der Prototyp wurde für die Erprobung zunächst mit Auftragsdaten der Produktgruppe »Seilwinden« erstellt. Die implementierte Lösung führt eine semantische Suche auf den Freitextfeldern ein, indem diese durch Embedding-Modelle encodiert werden. Die erstellten Embeddings der bereits realisierten Kundenlösungen werden dazu in einer Vektordatenbank gespeichert und für das spätere Retrieval herangezogen. Ziel war ein zuverlässiges Auffinden relevanter Einträge; eine nachgelagerte Antwortgenerierung durch ein Sprachmodell (wie es in einer vollständigen RAG-Pipeline üblich wäre) war nicht erforderlich. Ein Ablauf-Schema ist in Abbildung 10 dargestellt.

Die im Retrieval-Schritt gefundenen, vergangenen und semantisch ähnlichen Aufträge werden den Mitarbeitenden als Referenz angezeigt. Dadurch werden Rechercheaufgaben im Angebotswesen und in der Auftragskonstruktion beschleunigt, die Wiederverwendung bewährter Lösungen erhöht und Risiken in der Preisfindung reduziert. Zudem kann das System perspektivisch Hinweise auf sinnvolle zukünftige Varianten liefern und so die Weiterentwicklung des Variantenbaukastens unterstützen. Nach eigenen Aussagen des Unternehmens soll der Prototyp zu einer Produktivanwendung ausgebaut werden.

Mehr Informationen zum Projekt:

www.ki-fortschrittszentrum.de/projekte

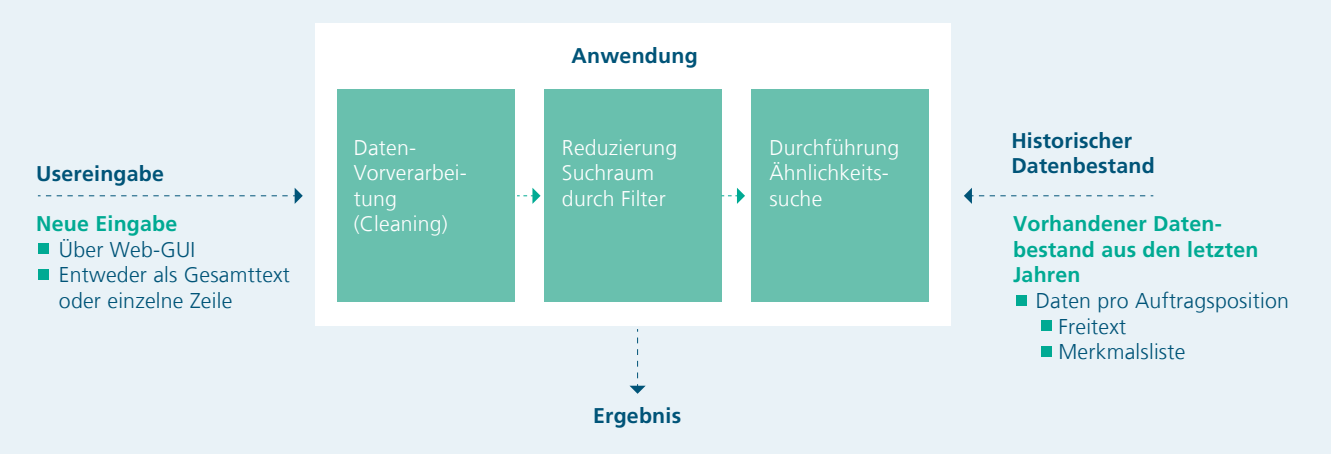


Abbildung 10: Schematische Darstellung der Verarbeitungspipeline des Prototyps. Die semantische Suche wird über eine User-Eingabe in der Web-Oberfläche des Tools angestoßen. Die Anfrage wird daraufhin aufbereitet, eine Eingrenzung des Suchraums auf den bestehenden Daten über weitere Merkmale vorgenommen und anschließend die Embeddings verglichen. Übereinstimmungen werden dem Nutzer angezeigt. Quelle: Eigene Darstellung.

3.6. Pilotprojekt: Campus Orientierungs-Assistent (COrA)

Das Pilotprojekt COrA für die Pädagogische Hochschule Ludwigsburg unterstützt im produktiven Testbetrieb die Mitarbeitenden des Prüfungsamts bei der Beantwortung von Standardfragen der Studierenden. Ziel war es, in einem klar abgegrenzten Anwendungsfall praktische Erfahrungen mit Konzeption, Umsetzung, Nutzung und Betrieb eines KI-Systems zu sammeln, das interne Abläufe unterstützt. Das beschriebene Szenario wurde durch das Prüfungsamt ausgewählt, da hier ein hoher Anteil ähnlicher Anfragen eingeht und die Antwortqualität stark von konsistenten, gültigen Formulierungen abhängt.

COrA wurde als Retrieval-Augmented-Generation-System umgesetzt. Der Kernansatz besteht darin, gültige Musterantworten als primäre Wissensbasis zu verwenden und daraus sprachlich passende, auf die konkrete Anfrage abgestimmte Antworten zu erzeugen. Um den Implementierungsaufwand gering zu halten, wurde der Prototyp bewusst ohne Schnittstellen zu bestehenden IT-Systemen der Hochschule realisiert. Damit adressiert die aktuelle Version ausschließlich allgemeine Fragen, für die keine personenbezogenen oder studiengangsspezifischen Daten erforderlich sind. Abbildung 11 zeigt die Benutzeroberfläche der Anwendung.

Technisch basiert COrA auf einem Standardkatalog aus Fragen und zugehörigen Musterantworten, der in einer Vektordatenbank abgelegt wird. Eingehende Anfragen werden in derselben Repräsentation verarbeitet und die fünf ähnlichsten Einträge werden abgerufen. Ein Sprachmodell prüft darauf aufbauend, ob eine hinreichend passende Referenzfrage vorliegt, und erstellt eine angepasste Antwort, die sich eng an der Musterantwort orientiert. Dies reduziert das Risiko von Halluzinationen, da keine zusätzlichen Aussagen ergänzt werden. Den Mitarbeitenden werden sowohl die Original-Musterantwort als Referenz als auch die generierte Variante bereitgestellt. Anpassungen können direkt vorgenommen und anschließend als vorformulierte Antwort-E-Mail genutzt werden. Der Prototyp stand dem Prüfungsamt über mehrere Wochen zur Erprobung zur Verfügung. Das Feedback nach dem Testlauf fiel durchweg positiv aus und Mitarbeitende äußerten den Wunsch, die Anwendung weiter nutzen zu wollen.

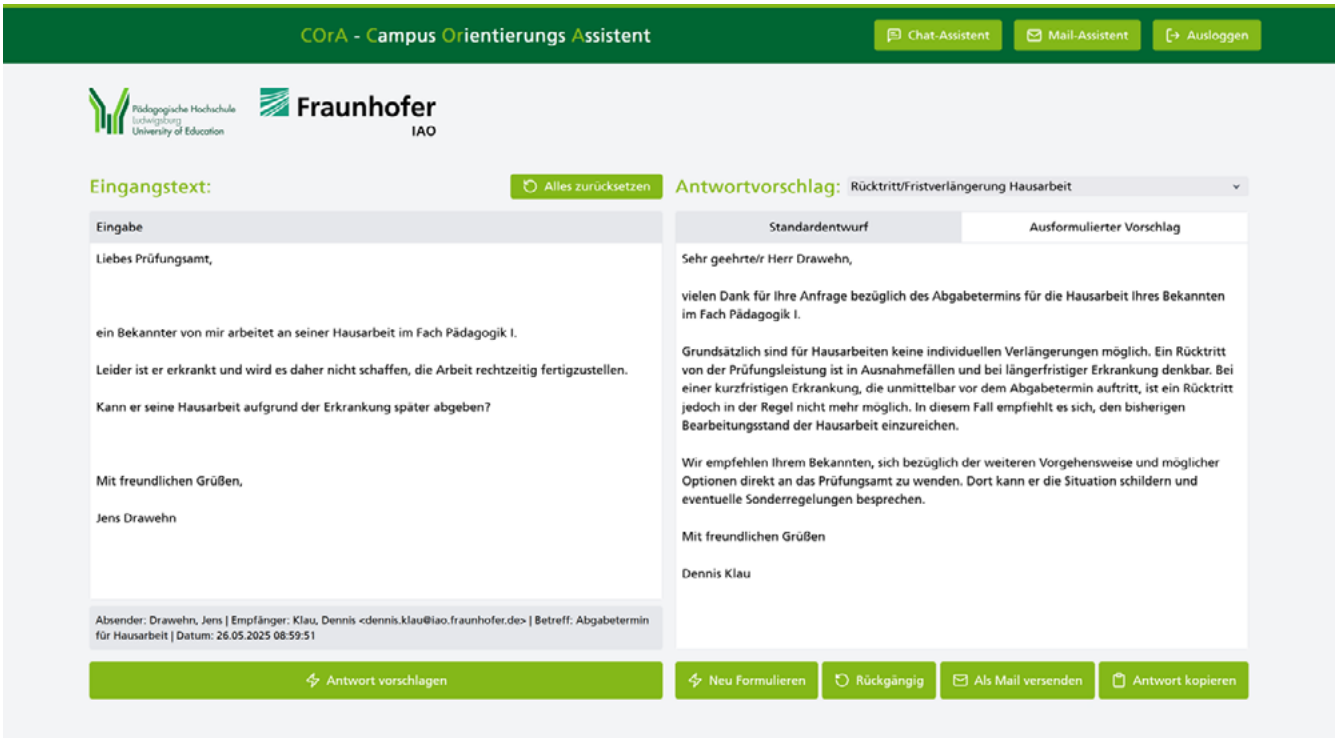


Abbildung 11: Nutzeroberfläche des Campus Orientierungs-Assistenten (COrA). Aus einer eingehenden E-Mail (links) wird die relevante Frage aus dem hinterlegten Fragenkatalog ermittelt und zusammen mit einem ausformulierten Antwortvorschlag angezeigt (rechts). Der Vorschlag kann dann entweder so übernommen oder angepasst werden. Quelle: Eigene Darstellung.

4. Untersuchung von RAG-Produktivsystemen

Neben den im Kapitel 3 dargestellten RAG-Proof of Concepts (PoCs), vor allem aus dem KI-Fortschrittszentrum, wurden im Rahmen der Studie durch etwa einstündige, leitfadengestützte Interviews fünf konkrete, produktiv umgesetzte RAG-Anwendungsfälle aufgenommen. Annahmen auf Basis von Literaturrecherche und Erfahrungen aus den vorher beschriebenen PoCs werden durch diese Interviews validiert und um Aspekte aus der unternehmerischen Praxis erweitert. Das folgende Kapitel fasst als Ergebnis der Interviews Gemeinsamkeiten und Unterschiede der untersuchten RAG-Projekte und -Systeme zusammen. Die Ergebnisse sind nicht repräsentativ, stärken jedoch die Handlungsempfehlungen im nächsten Kapitel.

Untersucht und an dieser Stelle anonymisiert ausgewertet wurden RAG-Anwendungsfälle im Forschungsumfeld, bei einer Krankenkasse, in der Versicherungs- und der Automobilbranche. In vier von fünf Fällen handelt es sich um größere Organisationen mit mehreren tausend Mitarbeitenden. Die interviewten Personen sind jeweils in mindestens ein RAG-Projekt involviert – manche davon auch als Entwickelnde an der technischen Umsetzung beteiligt. Alle sind in Rollen mit starkem IT-Bezug tätig, manche bringen darin auch einen wirtschaftswissenschaftlichen bzw. kaufmännischen Hintergrund ein.

4.1. Interne Anwendungsfälle zur Wissenssuche

»Chatbots, die auf interne Daten zugreifen, gehören zu den derzeit wichtigsten KI-Anwendungsfällen.«

Diese These wurde zusammen mit drei anderen Thesen jeweils am Ende eines Interviews präsentiert. Alle interviewten Personen bestätigten diese Aussage.

RAG-Anwendungsfälle sind sehr gefragt. Die Nachfrage nach niederschwelligen und einfach zu bedienenden Such- und Aufbereitungslösungen von intern vorhandenem Wissen in internen Dokumenten oder Wikis beschrieben alle interviewten Personen als sehr hoch. Fachbereiche und deren Mitarbeitende wünschen sich konkret spezifische Chatbots, die ihre Arbeit unterstützen. Als Generative KI-Anwendung steht in den Organisationen allen Mitarbeitenden bereits ein Alltags-Chatbot (meist eine semi-interne, cloud-basierte LLM-Variante einer der großen Modellanbieter) zur Verfügung. Dieser ist allerdings oft (noch) nicht an interne Wissensspeicher angebunden und kann somit nur bedingt bei firmenspezifischen Fragen weiterhelfen.

Am Anfang geht es darum, RAG auszuprobieren. Zielsetzung für die Umsetzung erster RAG-Systeme war in den meisten Organisationen, den RAG-Ansatz zu erproben und Erfahrungen damit zu sammeln. Ein RAG-Projekt wird immer mit einem Fachbereich durchgeführt und hat einen stark eingegrenzten Anwendungsfall adressiert. Dadurch blieb der Aufwand für Datenaufbereitung und Evaluation handhabbar.

Konkretes Ziel des RAG-Systems war in den Unternehmen immer auch ein Effizienzgewinn für die Zielgruppe – oftmals im Bereich der Sachbearbeitung. Für die Umsetzung weiterer RAG-Systeme ist häufig eine Nutzwertanalyse oder die Darstellung des Business Case etabliert. Die Vorhersage und Messung der Effekte von spezifischen Chatbots stellen eine Herausforderung dar. Die Experten weisen darauf hin, dass für andere Anwendungsfälle bzw. in klassischen Machine-Learning-Projekten, die üblicherweise strukturierte Daten nutzen und Dunkelverarbeitungsprozesse ermöglichen, wesentlich größere Einsparungen zu erwarten sind.

Alle untersuchten Anwendungsfälle sind intern, sie stehen also Mitarbeitenden zur Verfügung und richten sich nicht direkt an Kunden oder Partner. Keine der interviewten Personen konnte einen in ihrer Organisation fest eingeplanten, nach außen gerichteten produktiven Anwendungsfall benennen. Lediglich ein Pilot mit einem Dienstleister wurde in diesem Bereich benannt. Ein genannter Grund für die Konzentration auf interne RAG-Systeme nach dem Prinzip »*Chat mit den eigenen Daten*« sind Haftungs- und Image-Risiken, die bei der Kommunikation von Chatbots nach außen entstehen können.

Auffällig ist, dass derzeit **fast alle Unternehmen RAG-Systeme für weitere interne Anwendungsfälle** planen bzw. sich diese bereits in der Umsetzung befinden. Die Grundarchitektur bleibt dabei gleich und es wird für die nächsten ein bis zwei Jahre in diesem Zusammenhang von einer »Fertigungsstraße« gesprochen, in der das Backlog abgearbeitet wird. Der Betrieb und die Weiterentwicklung einer Vielzahl von RAG-Chatbots wird als zukünftige Aufgabe genannt. Perspektivisch stellen sich alle Befragten auch ein Portal für alle Chatbots der Organisation vor. Die Datenbasis bleibt dabei, so sehen es die interviewten Personen, getrennt. Die Rolle eines Nutzers schränkt dabei ein, welche Chatbots jeweils zur Verfügung stehen.

4.2. Projektorganisation

Alle untersuchten **Anwendungsfälle sind erst seit 2025 produktiv oder noch in der Entwicklung.** Ihre **Umsetzung hat im Schnitt ein Jahr gedauert**. Pausen gingen häufig mit Personalengpässen einher oder damit, dass Strukturen für weitere sehr ähnliche Anwendungsfälle geschaffen wurden. Auch Nutzertests und Freigaben, zum Beispiel durch den Betriebsrat oder Datenschützer, haben viel Zeit in Anspruch genommen.

Gestartet wird oft mit sehr einfachen Anwendungsfällen und einer sehr eingeschränkten Datenbasis, die dann schrittweise, selbst im Betrieb, erweitert wird. Es gibt in einzelnen Organisationen schon gute Erfahrungen mit unterschiedlichen Arten von Wissen, welches durch RAG besser auffindbar wird.

Initiiert wurden die Projekte häufig aufgrund des generellen Hypes um Generative KI und RAG. Erste Projekte waren dabei häufig jene, die aus einem Fachbereich getrieben und finanziert wurden.

Ein agiles Vorgehen mit Fokus auf schneller Auslieferung stand bei der ersten Umsetzung eines RAG-Chatbots im Vordergrund. Gestartet wurde immer mit einem Proof of Concept (PoC), gefolgt von einer Pilotumsetzung (oder Minimal Viable Product MVP) und schließlich der Überführung in ein Produktivsystem.

PoCs wurden entweder in der Fachabteilung selbst, dort teilweise mit Dienstleistern, umgesetzt oder ohne weitere Finanzierung mit der eigenen IT-Abteilung. Die meisten Organisationen sehen **für Piloten und Produktivsysteme eine Finanzierung durch die Fachabteilung** vor. Es gibt aber auch Modelle, in denen ein allgemeines strategisches Budget für Projekte im Kontext Generativer KI vorhanden war, worüber die RAG-Vorhaben umgesetzt werden konnten.

Das Projektteam bestand bei den untersuchten Anwendungsfällen aus **drei bis zehn Personen,** wobei größere Teams eher die Ausnahme darstellen. Es war für das Produktivsystem **immer in der eigenen IT-Abteilung angesiedelt** mit zwei bis drei Entwicklern, ggf. einem Data Scientist und meistens mit einem oder zwei Domänen-Experten aus dem Fachbereich. In einigen Fällen wurden im Laufe des Projekts **neue Teams in der IT** zusammengestellt oder sogar neue Teams für den Betrieb gebildet. Vor allem in der Anfangsphase wurde häufig nicht Vollzeit an den RAG-Anwendungsfällen gearbeitet und Knowhow dadurch nur langsam aufgebaut.

Alle untersuchten Projekte zur produktiven Umsetzung wurden **mit internen Ressourcen bearbeitet.** Dienstleister wurden zwar betrachtet, kamen aber letztendlich nur im Rahmen von PoCs zum Zug. Teilweise lag das daran, dass der eigene PoC bessere Ergebnisse lieferte. Hauptgrund für die Entscheidung war für die meisten aber, dass intern RAG-Knowhow aufgebaut werden sollte.

Alle RAG-Systeme wurden auf einer Cloud-Plattform implementiert, da diese in den Unternehmen schon für Datenhaltung und andere Services zur Verfügung steht. Eine **direkte Unterstützung durch den Betreiber der Cloud-Plattform** (zumeist Microsoft) wurde häufig in Anspruch genommen. Auch von einem Austausch innerhalb und außerhalb der eigenen Branche wurde berichtet.

4.3. Datengrundlage

»Identifikation, Beschaffung und Aufbereitung von Daten ist (noch) ein wesentlicher Teil von RAG-Projekten.«

Dieser These haben alle befragten Personen zugestimmt, wobei es in einem Unternehmen dazu, zum Zeitpunkt des Interviews, noch keine Erfahrungen gab. Die in datenfokussierten Bereichen wohlbekannte Aussage, dass Daten-Auf- und -Vorverarbeitung einen Großteil des Projektaufwands darstellen und somit auch in diesem Kontext wenig überraschend ist, wurde ebenfalls in einem Interview vorgebracht. Die Aufbereitung, z. B. die Texterkennung mit OCR, gehört beispielsweise noch nicht zu RAG an sich, wurde aber häufig in diesem Zusammenhang erwähnt.

Generell gibt es technische Möglichkeiten, die Datenaufbereitung zu unterstützen, die **inhaltliche Datenauswahl und -bereinigung ist jedoch vor allem Aufgabe des Fachbereichs,** in dem der RAG-Anwendungsfall genutzt werden soll. Dies führt oft zu Ernüchterung bei den

Fachabteilungen. Werden im PoC die Potenziale deutlich, motiviert das dazu, den notwendigen Aufwand zu betreiben. Eine andere Strategie ist, im ersten Schritt kleine, prompt-basierte Assistenten mit in Gänze übergebenen Dokumenten umzusetzen. Bei diesen steht dann im Vordergrund, einen Überblick über die verwendeten Daten zu behalten, sodass die Pflege und Verwaltung der Daten weiterhin möglich sind. Für das RAG-System werden in beiden Fällen **an einem neuen Ort Daten redaktionell aufgebaut und nicht direkt aus bestehenden Quellen bezogen.** In einigen Anwendungsfällen wurden auch direkt leicht verfügbare Daten z. B. aus Sharepoint verwendet. Dabei wurde festgestellt, dass auch in diesem Fall eine aufwendige Freigabe nicht ausbleiben kann. **In allen Projekten wird die Datengrundlage selbst im Produktivbetrieb noch erweitert und gepflegt.**

Alle umgesetzten Projekte verwenden **bislang nur textbasierte Daten,** eine Verarbeitung von weiteren Modalitäten wie Bildern, Videos oder Audios erfolgt noch nicht. Berichtet wird, dass auch bei Dokumenten wie Bedienungsanleitungen mit vielen Grafiken die Beschriftungen in bzw. um die Bilder bereits gute Ergebnisse liefern. Für die Datenbasis werden vor allem Dokumente verwendet, die unstrukturierte Daten enthalten. Für die zukünftige Anbindung von komplexeren Datenbanken werden agentische Systeme erwähnt, diese sind jedoch noch in keinem der untersuchten Beispiele im Einsatz.

Die Qualität der Daten bestimmt die Qualität der Ergebnisse des RAG-Systems. Entweder ist von vornherein eine gute Data Governance vorhanden oder die Datenqualität wird hauptsächlich manuell hergestellt oder schlechtere Ergebnisse werden hingenommen. Internetquellen sind teilweise vom Fachbereich gewünscht, aber aufgrund der schwierigen Handhabung bisher nicht integriert.

4.4. Technologie und Souveränität

Aus den Interviews wurde von einer anfänglichen Abwägung der unterschiedlichen Anpassungsvarianten generativer KI-Lösungen (siehe auch Kapitel 2.2) berichtet. Zwischen Training bzw. Finetuning von Modellen und In-Kontext-Wissensabrufen (RAG) erscheint ersteres allerdings derzeit weder wirtschaftlich noch zielführend und wurde von keiner Organisation weiterverfolgt. Eine Organisation gibt an, GraphRAG (siehe auch Kapitel 2.4.3) zukünftig in einem anderen Anwendungsfall auszuprobieren zu wollen.

Die These

»Retrieval-Augmented Generation (RAG) ist derzeit ein guter Ansatz, den wir sofern geeignet auch einsetzen.«

wird größtenteils bestätigt. Es wird vermutet, dass andere zukünftige Anwendungsfälle agentische Frameworks erfordern oder Wissensgraphen besser geeignet sind, um sehr große, verteilte Datenmengen und stärker strukturierte Daten einfließen zu lassen.

Alle Umsetzungen erfolgten auf einer kommerziellen Cloud-Plattform mit Beratung und Unterstützung des Cloud-Anbieters. Als Grund angegeben wurde unter anderem, dass alle benötigten Standardfunktionen z. B. zur Dokumentenvorverarbeitung verfügbar sind und man schnell zu Ergebnissen kommt. Außerdem ist die Nutzung von kommerziellen Clouds und deren Services in den Unternehmen Teil der Strategie.

Hinsichtlich der Souveränität der Lösungen herrscht Uneinigkeit. Einige der interviewten Personen sehen eine Art Anbieterbindung (Vendor Lock-In), die jedoch auch schon durch die Datenhaltung in der Cloud gegeben ist. Der Aufbau des RAG-Systems vor Ort (on Premise) oder bei anderen Cloud-Anbietern wird nach dem erfolgten Knowhow-Aufbau als einfach machbar angesehen. »*Open Source gibt es auch nicht geschenkt*« fasst ein Entwickler die Diskussion zusammen.

Allein aus Kostengründen werden günstigere Open-Source-Komponenten eingesetzt, die jedoch auch bei Cloud-Anbietern gehostet werden. Nur eine der fünf Organisationen nutzt nicht Microsoft Azure und setzt auch auf frei verfügbare (sog. Open Weight) Modelle bei einem anderen Cloud-Anbieter.

4.5. Akzeptanz und Evaluation

»Die Akzeptanz von Such- und Wissens-Anfragen an Chatbots ist besser bei RAG-Lösungen, die nachvollziehbare Ergebnisse liefern.«

Die These bezieht sich auf die Möglichkeit, mit einem RAG-System präzise aufzuzeigen, auf welchen spezifischen Quellen, bzw. welchen konkreten Sätzen darin, eine Ausgabe beruht. Damit unterscheiden sich spezifische RAG-Chatbots (noch) von Alltags-Chatbots. Alle interviewten Personen sind der Meinung, dass es sich in ihrem Anwendungsfall bei der **Quellenanzeige um einen wesentlichen Akzeptanz- und damit Erfolgsfaktor von RAG handelt.** Die Verlinkung der Quelldokumente, die nicht zwingend den Nutzenden angezeigt werden müsste, ist essenziell und steht für die meisten im Vordergrund. Einige beschreiben, dass Quelldokumente direkt geöffnet, an relevante Stelle gescrollt und die relevanten Abschnitte dort markiert werden sollen. Ein **Disclaimer**, also ein Texthinweis, der immer angezeigt wird und darauf hinweist, dass die KI-Lösung auch falsche Angaben erzeugen kann, wird als wichtiges Element erwähnt. Die Verantwortung für die Nutzung der Ergebnisse eines RAG-Chatbots bleibt damit bei den Nutzenden.

Alle interviewten Personen gaben an, dass **breite KI-Schulungsinitiativen und Workshops** zur allgemeinen KI-Anwendungen aus ihrer Sicht wichtig für die Akzeptanz des RAG-Systems sind. Geschult wird u.a. über die Verlässlichkeit der KI-Ergebnisse (siehe Disclaimer) – das führt bei den Mitarbeitenden oft zu Ernüchterung. Trotzdem überwiegt eine allgemeine KI-Begeisterung und die Schulung der Mitarbeitenden wird als Erfolgsfaktor für die Entwicklung, Einführung und Nutzung von KI-Systemen betrachtet. Ganz konkret entstehen über die Schulungen auch neue Anwendungsfälle.

In einem Fall benannt wurde die auffällig **hohe Akzeptanz und Nutzung des RAG-Chatbots durch das Management.** Es wurde die Hypothese aufgestellt, dass diese Zielgruppe sich zuvor selbst weniger tief mit den Daten auseinandergesetzt hat und nun mithilfe des RAG-Systems ganz andere Möglichkeiten dazu hat. Statt die Mitarbeitenden zu fragen, kann durch die neue Lösung der Manager selbst die benötigten Informationen für Entscheidungen zusammenstellen.

Die angezeigten Quellen schaffen bei RAG-Systemen nicht nur Akzeptanz, sondern auch die **Möglichkeit, Ergebnisse nachzuvollziehen und selbst zu validieren.** In der Evaluation der Anwendungsfälle sehen alle interviewten Personen noch Ausbaupotenzial, was vor allem daran liegt, dass diese Anwendungsfälle erst seit 2025 im Betrieb sind und meist noch weiter ausgebaut werden.

Evaluation durch Nutzerfeedback, z. B. durch Interviews, stand vor allem während der Entwicklung im Vordergrund. Im Betrieb sind ebenfalls Befragungen geplant, diese erfassen jedoch vor allem die subjektive Zufriedenheit der Nutzenden. Die systematische Erhebung von Leistungssteigerungen und Effizienzgewinnen durch die RAG-Systeme wird als schwierig erachtet. In einigen der untersuchten Anwendungsfälle wurde eine »Daumen hoch/runter«-Bewertung für Chatbot-Antworten umgesetzt. Bei der Auswertung technisch erhobener Nutzungswerte spielen Mitbestimmungsvorgaben und Datenschutz eine Rolle, weswegen diese teilweise noch nicht aktiv sind.

Weil die Lösungen teilweise noch sehr neu sind, werden die **Nutzungszahlen der untersuchten RAG-Systeme als ein Indikator für Akzeptanz und Wirksamkeit** als wenig aussagekräftig angesehen. Außerdem erlaubt die Nutzung noch keine Aussage über den Nutzen (die Wirksamkeit). Teilweise werden dazu umfangreiche Studien durchgeführt.

Dokumente mit Referenz-Fragen und zugehörigen -Antworten als ein sogenannter Gold-Standard-Datensatz liegt in allen Anwendungsfällen vor. Zuständig für dessen Erstellung ist der Fachbereich. Auf dieser Basis kommen häufig **durch Menschen durchgeführte Stichproben** zum Einsatz. LLM as a Judge, also ein Bewertungsverfahren, bei dem ein Sprachmodell bewertet, ob eine Antwort auf eine Testfrage mit der Referenzantwort übereinstimmt, wird bei den meisten Unternehmen erst bei weit entwickelten Piloten oder im Produktivbetrieb eingesetzt, wenn sich die Datenbasis weiterentwickelt und eine Pflege des Testdokuments nicht mehr tragbar ist.

4.6. Erkenntnisse und Ausblick

Die Analyse der Interviews fließt ein in die Handlungsempfehlungen durch das Fraunhofer IAO in Kapitel 5. An dieser Stelle wird angegeben, welche zusammenfassenden Erkenntnisse die interviewten Personen aus ihren Projekten angegeben haben.

Insgesamt waren alle interviewten Personen sehr zufrieden mit der Umsetzung des RAG-Anwendungsfalls und sehen den **internen Knowhow-Aufbau** als wichtig an. Häufig betont wird, dass der Anwendungsfall als Neben-Projekt in der IT gestartet ist und deshalb (zu) lange bis zur Inbetriebnahme brauchte. Dedizierte Ressourcen, insbesondere Personal, wären von Anfang an wünschenswert, bemerken die Macher. **Bestehende Organisationseinheiten müssen gegebenenfalls erweitert oder ergänzt** werden, das kostet Zeit.

Die Umsetzung des ersten RAG-Anwendungsfalls hat dazu geführt, eine technische Basis zu schaffen, die zukünftig für weitere Anwendungsfälle genutzt wird (Blaupause). Zusammenfassend bemerkte eine interviewte Person, dass man **früher Grundlagen, Templates und Strukturen aufbauen** sollte, da die nächsten zwei Jahre **sicherlich noch einige ähnliche Anwendungsfälle nachkommen.**

Der zukünftige **Betrieb und die Weiterentwicklung** der implementierten RAG-Systeme sind für einige Organisationen noch nicht gelöst. Fast immer ist das **Hinzufügen weiterer Datenquellen** geplant. In diesem Zusammenhang werden ebenso **agentische Ansätze** erwähnt, die erprobt werden sollen. Auch **Erweiterungen des RAG-Systems,** wie beispielsweise um Query-Expansion-Komponenten, steht für einige noch auf dem Programm.

Weitere Einzelaussagen betreffen außerdem das Thema **Souveränität**. Dieses sei den Auftraggebenden in den Organisationen häufig nicht bewusst. Es gibt wenig kritische Stimmen und häufig überzogene Erwartungen, berichtet eine interviewte Person. Auch wichtig zu beachten ist für die Anwendungsfälle, dass RAG-Anfragen länger dauern als normale Suchanfragen und höhere Kosten verursachen. Die finanzielle Bewertung des Nutzens eines RAG-Chatbots – vor allem langfristig – ist wegen dessen Natur und den genannten Herausforderungen in der Evaluation nicht trivial. Der Nutzen kann also nicht – wie in anderen Digitalisierungsprojekten – klar berechnet werden.

Die Frage »Brauchen wir an dieser Stelle einen spezifischen RAG-Chatbot?« muss öfter gestellt werden.

5. Handlungsempfehlungen für RAG-Projekte

In diesem Kapitel werden die Erkenntnisse aus den in Kapitel 3 beschriebenen, am Fraunhofer IAO erstellten RAG-Prototypen sowie die Untersuchung der RAG-Produktivsysteme in größeren Organisationen in Kapitel 4 zusammengeführt und Handlungsempfehlungen abgeleitet. Die Analyse umfasste zwei Workshops mit Fraunhofer-Wissenschaftlerinnen, sowie das Review der formulierten Handlungsempfehlungen durch alle Beteiligten.

Die Handlungsempfehlungen werden in vier Abschnitte gegliedert: Empfehlungen, die allgemein für KI-Projekte gelten und Teil der Unternehmensstrategie sein sollten (5.1), Empfehlungen für das erste RAG-Projekt (5.2), Empfehlungen für Operationalisierung und Betrieb von RAG-Systemen (5.3) sowie Empfehlungen für Zukunftsthemen und Weiterentwicklung (5.4). Die folgende Tabelle gibt einen kurzen Überblick über die Zielsetzungen in jeder einzelnen Phase und fasst Kernaktionen zusammen.

Tabelle 5: Übersicht Handlungsempfehlungen für RAG-Projekte.

Abschnitt/Phase	Ziel	Kernaktionen
5.1 KI-Projekte allgemein	Grundlagen und Akzeptanz schaffen	Datenstrategie und Data Governance, Berührungspunkte mit KI schaffen/Mitarbeitende schulen
5.2 Erstes RAG-Projekt	Schneller, risikoarmer Einstieg	Interner Chatbot als PoC mit eigenem Datenspeicher, Quellenanzeige, Disclaimer und Risikoabschätzung
5.3 Betrieb RAG-Systeme	Stabiler Betrieb und Skalierung	Technische Blaupause mit kontinuierlicher Evaluation, Kosten- und Betriebsmodell
5.4 Zukunftsthemen	Qualität steigern	Datenqualitäts-Feedback, Datenranking und Verwendung von bewerteten Ergebnissen als zusätzliche Quelle

5.1. Allgemeine Empfehlungen bezüglich KI

Die Empfehlungen in diesem Abschnitt gelten für alle KI-Projekte und sind nicht spezifisch für RAG.

Schaffen Sie Berührungspunkte mit KI. Ermöglichen Sie zum Beispiel Ihren Mitarbeitenden die Nutzung eines Alltags-Chatbots. Dadurch wird verhindert, dass interne Daten in öffentlichen KI-Werkzeugen verarbeitet werden (»Schatten-KI«). Mitarbeitende können dadurch Generative KI offiziell nutzen und lernen die Möglichkeiten, Herausforderungen und Grenzen dieser neuen Technologie und Interaktionsform kennen.

Schulen Sie Ihre Mitarbeitenden zum Thema KI. Unabhängig von RAG besteht ein großer Bedarf in den Organisationen, Akzeptanz für KI-Anwendungen zu schaffen und sichere KI-Nutzung zu ermöglichen. Darüber hinaus werden über KI-Schulungen oftmals KI-Anwendungsfälle identifiziert.

Kümmern Sie Sich um Datenstrategie und Data Governance. Daten sind die Grundlage aller KI-Projekte. RAG-Projekte, die interne Daten nutzen, sind umso einfacher und schneller umzusetzen, je besser die Datengrundlage ist.

5.2. Empfehlungen für das erste RAG-Projekt

Diese Empfehlungen richten sich an Organisationen, die RAG bisher noch nicht nutzen.

Setzen Sie einen kleinen RAG-Anwendungsfall intern um. Große Organisationen haben hier meist schon Erfahrungen und einen organisatorischen und technischen Rahmen für die Umsetzung von RAG-Anwendungsfällen für sich geschaffen. RAG wird auch zukünftig eine Rolle spielen – der Know-how-Aufbau lohnt sich. Ein interner Anwendungsfall birgt normalerweise keine Haftungs- oder Betriebsrisiken. Starten Sie mit einem kleinen Anwendungsfall, der später erweitert werden kann.

Ein interner KI-Chatbot zur Suche in spezifischen Daten ist ein geeigneter erster RAG-Anwendungsfall. Hierfür gibt es viele Beispiele und etablierte Frameworks. Nehmen Sie im PoC erstmal die Daten, die da sind. Zu beachten ist, dass der Nutzen solcher Chatbots im Sinne einer Effizienzsteigerung für Mitarbeitende schwer zu messen ist.

Stellen Sie Ressourcen für das RAG-Projekt bereit. Sie brauchen zwei bis drei Personen in der IT und ein bis zwei Personen im Fachbereich. Je mehr Zeit diese Personen aufwenden können, desto schneller geht das Projekt voran – die Umsetzung der untersuchten RAG-Produktivsysteme dauerte durch Freigabe, Nutzertests, jeweils ca. ein Jahr. Werden schnell mehrere RAG-Anwendungsfälle identifiziert, die umgesetzt werden sollen, kann organisatorisch die Schaffung eines neuen Teams und technologisch die Schaffung einer Blaupause notwendig sein.

Starten Sie mit einem Proof of Concept. Ein PoC für einen spezifischen Anwendungsfall kann schnell erstellt werden. Beachten Sie dabei vor allem die User Experience und die Anwendungstiefe. Die Ergebnisse des PoC liefern zahlreiche Erkenntnisse und motivieren die Fachabteilung.

Bauen Sie einen eigenen Datenspeicher für den RAG-PoC auf. Die Integration von Bestandsdaten kann schnell sehr aufwendig werden. Für den PoC können Daten an geeigneter Stelle neu und strukturiert aufgebaut werden. Mehr redaktioneller Aufwand verlängert auch die Umsetzungszeit für das Produktivsystem, jedoch verbessert es auch die Qualität der Ergebnisse.

Machen Sie eine Risikoabschätzung. Was kann schlimmstenfalls passieren, wenn Ergebnisse halluziniert werden? Leiten Sie daraus die notwendige Ergebnisqualität, benötigte Funktionen innerhalb des Systems und Regeln für die manuelle Überprüfung der Ergebnisse ab.

Die Anzeige und Verlinkung von Quellen sind unerlässlich. Die Möglichkeit, durch die Angabe von Quellen das Ergebnis einer Anfrage zu überprüfen, ist der größte Mehrwert von RAG. Damit können die Nutzenden das Ergebnis validieren oder sogar je nach Funktionalität der Lösung gegebenenfalls korrigieren. Für eine einfache und nutzerfreundliche Interaktion wird im Idealfall das jeweilige Quell-Dokument geöffnet, zur richtigen Stelle gescrollt und die Stelle hervorgehoben.

Verwenden Sie einen Disclaimer. Das Sprachmodell in RAG-Systemen kann falsche oder erfundene Angaben (Halluzinationen) erzeugen. Machen Sie klar, dass die Verantwortung weiterhin bei den Nutzenden liegt.

5.3. Empfehlungen für den Betrieb von RAG-Systemen

Ist ein erstes RAG-System schon implementiert, sind vor allem folgende Empfehlungen relevant:

Validieren Sie Ihre RAG-Ergebnisse kontinuierlich. Menschliche Stichproben sind eine einfache Möglichkeit, die schon während der Entwicklung zum Einsatz kommt. Bei LLM-as-a-Judge-Evaluationen bewertet ein Sprachmodell, ob das Ergebnis einer Frage zu einer vorab erfassten Musterantwort passt. Änderungen am System können dessen Verhalten stark beeinflussen und machen eine kontinuierliche Validierung notwendig. Evaluieren Sie darüber hinaus auch die Nutzendeninteraktion, z.B. über Feedback-Funktionen oder Verhaltenstracking. Beachten Sie dabei den Datenschutz.

Schaffen Sie eine technische Blaupause für weitere RAG-Anwendungsfälle. Beachten Sie, dass hier die Datenmenge relevant ist. Werden mehr Daten verarbeitet, sind im PoC oder Piloten verwendete Komponenten ggf. nicht mehr geeignet.

Behalten Sie die Betriebskosten im Blick. Neben der Implementierung erzeugt auch der Betrieb von RAG-Systemen zum Teil signifikante Kosten, beispielsweise für die Nutzung der Modelle in der Cloud, aber auch für weitere durchgeführte Vorverarbeitungsschritte und Services wie OCR.

Planen Sie den Betrieb der RAG-Systeme. Bei mehreren RAG-Systemen ist hierfür möglicherweise eine eigene Organisationseinheit notwendig, die die Pflege und Überwachung des Systems übernimmt.

Achten Sie auf Souveränität der RAG-Lösung. Machen Sie sich bei der Planung die direkten und indirekten Abhängigkeiten und mögliche Vendor Lock-Ins und deren Konsequenzen bewusst. Bewerten Sie diese Abhängigkeiten in regelmäßigen Abständen neu, um strategische, politische und marktabhängige Änderungen zu reflektieren.

5.4. Zukunftsthemen für RAG-Systeme

Empfehlungen für die Weiterentwicklung von RAG-Systemen sind:

Bauen Sie Funktionalitäten ein, um die Datengrundlage zu verbessern. Implementieren Sie ein Feedback der Nutzenden zur Datengrundlage, um diese zu verbessern und damit auch das RAG-System insgesamt zu optimieren. Schaffen Sie zum Beispiel Rückkopplungsmöglichkeiten bis zum Data Owner.

Verwenden Sie von den Nutzenden bewertete Ergebnisse als zusätzliche Datenquelle. Nutzende können über entsprechende zu schaffende Feedback-Mechanismen nicht nur als Konsumenten der Lösung, sondern auch als Verbesserer der Datengrundlage auftreten. Beispielsweise kann eine für eine Frage als gut bewertete Antwort als neue Musterantwort wiederverwendet werden. Entsprechende Qualitätssicherungsmaßnahmen sind für solche Ansätze Voraussetzung.

Standardisieren Sie Ihre Chatbots. Begegnen Sie der Entwicklung, dass in manchen Organisationen mehrere RAG-Chatbots entstehen. Schaffen Sie technologische und organisatorische Standards, um von Synergieeffekte zu profitieren und um die Nachteile einer nicht einheitlichen IT-Umgebung zu minimeren.

Ranken Sie Ihre Daten. Je mehr Daten in das RAG-System aufgenommen werden, desto wahrscheinlicher gibt es einen systematischen Qualitätsunterscheid aus verschiedenen Quellen: beispielsweise kann den Inhalten betriebseigener Dokumente wahrscheinlich mehr vertraut werden als dem Resultat einer Internetsuche. Hier wird eine Priorisierung der verschiedenen Daten für die Performanz des RAG-Systems sinnvoll.

Sammeln Sie Erfahrungen mit einem multimodalen RAG. Dies gilt, wenn neben Text andere Datenformate wie Bilder, Videos usw. ausgewertet werden sollen. Für RAG-Starter gilt: Starten Sie wenn möglich rein textbasiert. Auch im Fall vieler bildbasierter Inhalte ist die Wahrscheinlichkeit hoch, gute Ergebnisse allein durch die beschreibenden Texte und die im Bild erkannten Textelemente zu erzielen.

6. Fazit und Ausblick

Die Mittelstandsstudie 2025 (Harden, 2025) konstatiert, dass 72 Prozent der befragten Unternehmen KI nutzen, 32 Prozent davon systematisch. Die Frage nach den konkreten Anwendungsfällen bleibt hier und in anderen Studien jedoch offen. Die vom Fraunhofer IAO aufgestellte These, welche von den interviewten Personen bestätigt wurde, »*Chatbots, die auf interne Daten zugreifen gehören zu den derzeit wichtigsten KI-Anwendungsfällen.*« muss in Zukunft in größeren Stichproben untersucht werden, um Orientierungswissen zu schaffen.

Die Interviews in dieser Studie haben gezeigt, dass spezifische Chatbots ein häufiger Anwendungsfall sind und RAG dabei als Stand der Technik gilt. Es gibt in den Unternehmen produktive RAG-Systeme und noch viel Bedarf an der Umsetzung weiterer spezifischer Chatbots – einige Unternehmen sprechen von einer Implementierung am Fließband. Die Frage, ob es zukünftig viele RAG-Systeme pro Unternehmen gibt oder es zu einer Integration von Frontend oder sogar Backend kommen wird, ist noch offen.

Retrieval-Augmented Generation ist heute also für niemanden, der sich mit KI beschäftigt, ein unbekannter Begriff. Aber es ist auch heute noch kein abgeschlossenes Thema, sondern vielmehr ein Architektur- und Prozessprinzip, welches sich noch stark weiterentwickelt: Während *Naive RAG* bei einem Cloud-Anbieter mit wenigen Klicks umgesetzt werden kann, sind erweiterte RAG-Ansätze zu großen Teilen noch Forschungsthema und in der Praxis noch nicht relevant. Eine Pilotierung in geeigneten Anwendungsfällen hat jedoch großes Potenzial. Auch die Souveränität der Lösungen wird zunehmend eine Rolle spielen. Die untersuchten Unternehmen, die bereits RAG-Systeme einsetzen, haben auch einen großen Backlog an Projekten, die noch umgesetzt werden könnten. Im Vordergrund steht bei der Priorisierung nicht mehr, dass Erfahrungen mit KI und RAG gesammelt werden sollen, sondern der wirtschaftliche Mehrwert jedes einzelnen Chatbots. Die Herausforderungen dabei sind nicht in erster Linie technisch: Fachliche Datenauswahl, -bereinigung, -aufbereitung und -anbindung nehmen bei der Umsetzung eines spezifischen Chatbots häufig noch viel Zeit in Anspruch. Hinzu kommen Datenschutz und Compliance sowie generelle Akzeptanz-Themen.

Der Betrieb von RAG-Systemen wird in Zukunft zusätzliche Personal-Ressourcen binden. Und je größer die Nutzungszahlen werden, desto höher sind die Kosten, die durch die Anfragen an die LLMs entstehen. Die Bedeutung von selbst betriebenen Open-Weight-Modellen wird dadurch zunehmen. Mit dem Aufkommen von KI-Agenten wandelt sich ebenfalls das Paradigma des Informationszugriffs. Während bisherige RAG-Ansätze nach dem Push-Prinzip Informationen vorab zusammengestellt haben, geht Agentische KI nach dem Pull-Prinzip vor: das System entscheidet selbstständig, welche Informationen fehlen und ruft diese dynamisch ab (Härter, 2025). Agentische RAG stellt hier eine zukünftig vielversprechende Lösung dar, wenn Flexibilität und Autonomie in der Aufgabenbewältigung sowie eine dynamische Einbindung mehrerer unterschiedlicher Datenquellen benötigt werden. Mehr Informationen zu Agentischer KI und deren Einsatzmöglichkeiten liefert die Studie von (Büllesfeld et al., 2026).

Wenn es in dieser Studie nur ein Fazit geben darf, dann wäre es für die Autorin und den Autor dieses:

»Zukunftsfähig sind nicht jene Unternehmen, die heute die meisten oder besten RAG-Anwendungsfälle umgesetzt haben, sondern jene, die durch die Umsetzung einer Datenstrategie ihren gesamten Datenbestand KI-ready machen.«

7. Interview-Leitfaden

1. **Informationen zur Person und zu der Organisation der Interviewpartnerin/ des Interviewpartners**

2. **RAG-Lösung**
Bei mehreren RAG-Anwendungsfällen den wichtigsten Anwendungsfall auswählen.

- Use Case
  - Was ist der RAG-Anwendungsfall in Ihrem Unternehmen, über den wir heute sprechen?
  - Warum wurde dieser Anwendungsfall gewählt?
  - Welche Zielgruppen adressiert die Lösung?
  - Für welche anderen Anwendungsfälle steht dieser Anwendungsfall beispielhaft?
  - Was wollte man (ursprünglich) erreichen? PoC oder Produktivsystem?
  - Welche Kriterien spielen bei der Auswahl und Prüfung von Anwendungsfällen in Ihrer Organisation eine Rolle?

- Projekt
  - Wer war der interne Sponsor?
  - Welche Voraussetzungen/Vorarbeiten waren vorhanden? Z.B. bereits bestehende SW-Lösungen.
  - Team: Wer war wie beteiligt? Wie viele Leute insgesamt? Welcher Kenntnisstand? Welche Kompetenzen? Dienstleister ja/nein, wie?
  - Zeitschiene: Wie lange hat das Projekt insgesamt gedauert (von der Idee bis zur Umsetzung)? Welche Phasen? Gab es Pausen oder Blocker?

- Datengrundlage
  - Welche Daten(-Quellen) werden für die Anwendung verwendet?
  - Wurden die Daten aufbereitet? Wie? Warum?
  - Wie war der Prozess zur Auswahl der Daten?
  - Könnte/sollte Ihrer Meinung nach die Datengrundlage erweitert werden?
  - Wie war die Zeitschiene von der Auswahl bis zur Bereitstellung der Daten?

- Technologie
  - Warum fiel die Entscheidung für einen RAG-Ansatz? Waren alternative Ideen im Gespräch?
  - Welche Frameworks werden genutzt? Warum?
  - War die Art der Technologie (Open-Source vs. proprietär) ein Kriterium?
  - Souveränität: Handelt es sich um eine öffentliche oder private Cloud oder sogar eine On-premise-Lösung?

- Validierung
  - Wie wird der Erfolg bzw. die Qualität der Ergebnisse gemessen?
  - Wie sehen Tests in diesem Zusammenhang aus?

- Learnings und Ausblick
  - Was würden Sie heute im Projekt ggf. anders machen (Beteiligte, Technologien)?
  - Was waren die größten Herausforderungen im Projekt?
  - Welche Folgeaktivitäten ergaben sich aus dem Projekt?
  - Was ist ggf. der nächste (mit RAG) umzusetzende Anwendungsfall?
  - Welche Auswirkungen hat der Anwendungsfall intern und ggf. extern? Welche Auswirkungen halten Sie in Zukunft für wahrscheinlich?

- Akzeptanz der Lösung
  - Wie wird die RAG-Lösung von der Zielgruppe eingesetzt?
  - Entspricht die Nutzungshäufigkeit den Erwartungen?
  - Wie ist die Wahrnehmung der Lösung? Was sagen die Nutzenden?
  - Wer promotet die Lösung? Wer ist kritisch? Warum?

3. **Validierung von Thesen**
Nun stellen wir Ihnen eine Reihe von Thesen zu RAG vor und möchten Sie bitten, uns mitzuteilen, ob Sie diesen zustimmen.

Chatbots, die auf interne Daten zugreifen, gehören zu den derzeit wichtigsten KI-Anwendungsfällen.
Stimmen Sie dieser These zu?
Nein Teilweise Ja

Retrieval-Augmented Generation ist derzeit ein guter Ansatz, den wir sofern geeignet auch einsetzen.
Stimmen Sie dieser These zu?
Nein Teilweise Ja

Identifikation, Beschaffung und Aufbereitung von Daten ist (noch) ein wesentlicher Teil von RAG-Projekten.
Stimmen Sie dieser Aussage zu?
Nein Teilweise Ja

Die Akzeptanz von Such- und Wissens-Anfragen an Chatbots ist besser bei RAG-Lösungen, die nachvollziehbare Ergebnisse liefern.
Stimmen Sie dieser Aussage zu?
Nein Teilweise Ja

8. Literaturverzeichnis

Abootorabi, M. M., Zobeiri, A., Dehghani, M., Mohammadkhani, M., Mohammadi, B., Ghahroodi, O., . . . Asgari, E. (2025). Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. *Findings of the Association for Computational Linguistics: ACL 2025* (S. 16776-16809). Vienna, Austria: Association for Computational Linguistics.

ARD/ZDF-Forschungskommission. (2025). *ARD/ZDF-Medienstudie 2025 (Basispräsentation).* https://www.ard-zdf-medienstudie.de/.

Büllesfeld, E., Falkner, J., Kintz, M., Klau, D., Proissl, C., Schuller, A., . . . Voß, A. (2026). *KI-Agenten verstehen und anwenden: Einordnung und Einsatz von KI-Agenten.* Fraunhofer Heilbronn Forschungs- und Innovationszentren HNFIZ - Forschungs- und Innovationszentrum für Hybride Künstliche Intelligenz: Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO und Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. doi:http://dx.doi.org/10.24406/publica-6838

Carpineto, C., & Romano, G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys* (S. 50). New York, NY, USA: Association for Computing Machinery.

Dang, V., Bendersky, M., & Croft, W. B. (2013). Two-Stage Learning to Rank for Information Retrieval. In *Advances in Information Retrieval* (S. 423-434). Berlin, Heidelberg: Springer Berlin Heidelberg.

DeepMind, G. (2025). *Gemini 3 Pro – Model Card.* Von Google DeepMind: https://deepmind.google/models/model-cards/gemini-3-pro abgerufen

Drawehn, J., & Pohl, V. (2023). *Einsatz von KI mit Fokus Kundenkommunikation.* Stuttgart: Fraunhofer IAO. doi:http://dx.doi.org/10.24406/publica-1704

Faloutsos, C., & Oard, D. W. (1995). A survey of information retrieval and filtering methods. *Technical Report CS-TR-3514, University of Maryland*.

Fraga, N. (2024). Challenging LLMs Beyond Information Retrieval: Reasoning Degradation with Long Context Windows. *Preprints*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., . . . Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey.* abs/2312.10997: ArXiv.

Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., . . . Tang, J. (2024). Retrieval-Augmented Generation with Graphs (GraphRAG). *ArXiv*, abs/2501.00309.

Harden, L. H. (2025). *Die große Mittelstandsstudie 2025.* https://www.ingutergesellschaft.org/mittelstandsstudie#ki : ZEIT und Stiftung „in guter Gesellschaft".

Härter, H. (16. 12 2025). *Warum eine KI läuft, aber noch nicht liefert*. Von BigData Insider: https://www.bigdata-insider.de/warum-eine-ki-laeuft-aber-noch-nicht-liefert-a-7a3db5b-560f6373a9247125102ebeee7 abgerufen

Hong, K., Troynikov, A., & Huber, J. (2025). *Context Rot: How Increasing Input Tokens Impacts LLM Performance.* https://research.trychroma.com/context-rot: Chroma.

Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., & Zhao, L. (2025). GRAG: Graph Retrieval-Augmented Generation. *Findings of the Association for Computational Linguistics: NAACL 2025* (S. 4145–4157). Albuquerque, New Mexico: Association for Computational Linguistics.

Kim, K., & Lee, J.-Y. (2024). RE-RAG: Improving Open-Domain QA Performance and Interpretability with Relevance Estimator in Retrieval-Augmented Generation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (S. 22149-22161). Miami, Florida, USA: Association for Computational Linguistics.

Kintz, M., Beinhauer, W., Bienzeisler, B., Drawehn, J., Dworschak, B., Engelbach, M., . . . Wulf, J. (2024). *Potenziale Generativer KI für den Mittelstand: Wie große KI-Modelle die Arbeitswelt verändern.* Stuttgart: Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO). doi:10.24406/publica-2246

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (S. 16). Red Hook, NY, USA: Curran Associates Inc.

Loosen, W. (2016). Das Leitfadeninterview – eine unterschätzte Methode. In S. Averbeck-Lietz, & M. Meyen, *Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft* (S. 139–155). Wiesbaden: Springer Nature.

Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query Rewriting in Retrieval-Augmented Large Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (S. 5303-5315). Singapore: Association for Computational Linguistics.

Maslej, N., Loredana, Perrault, R., Gil, Y., Parli, V., Kariuki, N., . . . Walsh, T. (2025). *Artificial Intelligence Index Report 2025.* arXiv: arxiv.org/abs/2504.07139.

Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). *A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions.* arXiv preprint arXiv:2507.18910.

OpenAI. (13. 11 2023). *A Survey of Techniques for Maximizing LLM Performance.* Von YouTube: https://www.youtube.com/watch?v=ahnGLM-RC1Y abgerufen

Riezler, S., & Liu, Y. (2010). Query Rewriting Using Monolingual Statistical Machine Translation. *Computational Linguistics*, 569-582.

Schaaf, N., Wiedenroth, S. J., & Wagner, P. (2021). *Erklärbare KI in der Praxis: Anwendungsorientierte Evaluation von XAI-Verfahren.* Stuttgart: Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA. doi:10.24406/publica-fhg-300845

Singh, A. P., Jamdar, A., & Kaul, P. (2025). Agentic Retrieval-Augmented Generation: Advancing AI-Driven Information Retrieval and Processing. *International Journal of Computer Trends and Technology (IJCTT)*, 91-97.

Weber, I., Linka, H., Mertens, D., Muryshkin, T., Opgenoorth, H., & Langer, S. (2024). FhGenie: a custom, confidentiality-preserving chat AI for corporate and scientific use. *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)* (S. 26-31). IEEE.

Über Fraunhofer

Fraunhofer-Gesellschaft

Forschen für die Praxis ist die zentrale Aufgabe der Fraunhofer-Gesellschaft. Die 1949 gegründete Forschungsorganisation betreibt anwendungsorientierte Forschung zum Nutzen der Wirtschaft und zum Vorteil der Gesellschaft. Vertragspartner und Auftraggeber sind Industrie- und Dienstleistungsunternehmen sowie die öffentliche Hand.

Die Fraunhofer-Gesellschaft betreibt in Deutschland derzeit 74 Institute und Forschungseinrichtungen. Mehr als 28000 Mitarbeiterinnen und Mitarbeiter, überwiegend mit natur- oder ingenieurwissenschaftlicher Ausbildung, erarbeiten das jährliche Forschungsvolumen von mehr als 2,8 Milliarden Euro. Davon entfallen mehr als 2,3 Milliarden Euro auf den Leistungsbereich Vertragsforschung. Rund 70 Prozent dieses Leistungsbereichs erwirtschaftet die Fraunhofer-Gesellschaft mit Aufträgen aus der Industrie und mit öffentlich finanzierten Forschungsprojekten. Rund 30 Prozent werden von Bund und Ländern als Grundfinanzierung beigesteuert, damit die Institute Problemlösungen entwickeln können, die erst in fünf oder zehn Jahren für Wirtschaft und Gesellschaft aktuell werden.

Internationale Kooperationen mit exzellenten Forschungspartnern und innovativen Unternehmen weltweit sorgen für einen direkten Zugang zu den wichtigsten gegenwärtigen und zukünftigen Wissenschafts- und Wirtschaftsräumen.

Mit ihrer klaren Ausrichtung auf die angewandte Forschung und ihrer Fokussierung auf zukunftsrelevante Schlüsseltechnologien spielt die Fraunhofer-Gesellschaft eine zentrale Rolle im Innovationsprozess Deutschlands und Europas. Die Wirkung der angewandten Forschung geht über den direkten Nutzen für die Kund*innen hinaus: Mit ihrer Forschungs- und Entwicklungsarbeit tragen die Fraunhofer-Institute zur Wettbewerbsfähigkeit der Region, Deutschlands und Europas bei. Sie fördern Innovationen, stärken die technologische Leistungsfähigkeit, verbessern die Akzeptanz moderner Technik und sorgen für die Aus- und Weiterbildung des dringend benötigten wissenschaftlich-technischen Nachwuchses.

Ihren Mitarbeiterinnen und Mitarbeitern bietet die Fraunhofer-Gesellschaft die Möglichkeit zur fachlichen und persönlichen Entwicklung für anspruchsvolle Positionen in ihren Instituten, an Hochschulen, in Wirtschaft und Gesellschaft. Studierenden eröffnen sich aufgrund der praxisnahen Ausbildung und Erfahrung an Fraunhofer-Instituten hervorragende Einstiegs- und Entwicklungschancen in Unternehmen.

Namensgeber der als gemeinnützig anerkannten Fraunhofer-Gesellschaft ist der Münchner Gelehrte Joseph von Fraunhofer (1787–1826). Er war als Forscher, Erfinder und Unternehmer gleichermaßen erfolgreich.

Fraunhofer IAO

Mensch und Technik in der digitalen Arbeitswelt, Wirtschaft und Gesellschaft
Digitale Technologien verändern unsere Arbeitswelt und haben tiefgreifende Auswirkungen auf Wirtschaft und Gesellschaft. Lang etablierte Methoden und Prozesse werden in kurzer Zeit modernisiert und revolutioniert. Das Fraunhofer IAO kooperiert eng mit dem Partnerinstitut IAT der Universität Stuttgart und entwickelt gemeinsam mit Unternehmen, Institutionen und Einrichtungen der öffentlichen Hand wirksame Strategien, Geschäftsmodelle und Lösungen für die digitale Transformation.

Die digitale Transformation und neue IT-Technologien eröffnen für Unternehmen viele Chancen: innovative Produktangebote für neue Zielgruppen, bessere und kostengünstigere Prozesse, eine »intelligentere« Kundenkommunikation und höhere Automatisierung. Dafür kommen innovative, vernetzte IT-Lösungen auf Basis von Big Data, Künstlicher Intelligenz, Cloud und Internetplattformen zum Einsatz.

Die richtige Strategie und IT sind eine wesentliche Grundlage für den Erfolg und die Wettbewerbsfähigkeit von Unternehmen. Voraussetzung für erfolgreiche Anwendungen ist ein klarer Nutzen für das Unternehmen, seine Kund*innen und seine Partner.

Unsere Leistungen basieren auf fundierter Technologie- und Marktkenntnis sowie branchenübergreifenden Erfahrungen. Durch den Einsatz unserer praxiserprobten Methoden und erfahrenen Mitarbeitenden sichern wir den Projekterfolg. Unser Fraunhofer-Netzwerk ermöglicht uns den Zugriff auf ein umfassendes Kompetenzspektrum.

Das Fraunhofer IAO und das IAT der Universität Stuttgart beschäftigen gemeinsam mehr als 650 Mitarbeitende und verfügen über rund 15 000 Quadratmeter Büroflächen, Demonstrationszentren sowie Entwicklungs- und Testlabors.

Der Forschungsbereich Digital Business

Der Forschungsbereich »Digital Business« unterstützt Unternehmen dabei, die Chancen der Digitalisierung durch zukunftsfähige Strategien und neue IT-Lösungen zu nutzen.

Die Fokusthemen des Forschungsbereichs »Digital Business« sind:

- Digitalisierungsstrategien und -lösungen, u. a. Innovationsnetzwerke, Zukunftsszenarien, Benchmarks, New Business Radar, Data-Science-Seminare
- Datenbasierte Automatisierungslösungen und Künstliche Intelligenz, u. a. Prozessautomatisierung, digitale Assistenten, maschinelle Lernverfahren, wissensbasierte Arbeitsplätze, RAG-Chatbots, Big-Data-Lösungen
- Cloudbasierte Plattformen und Geschäftsmodelle, u. a. webbasierte Services, Smart Products und Services, Serviceorientierte Architekturen, Ökosysteme und Wertschöpfungsketten, Sicherheit und Datenschutz, Datenmanagement sowie Data Sharing, Innovation und Kooperation
- Integrierte IT-Systeme für das Internet of Things (IoT), u. a. IoT-Lösungen, Echtzeit-Datenplattformen, Analysesysteme, Steuerungslösungen, autonome Systeme, Smart City
- Unternehmenssoftware und Prozessoptimierung, u. a. Enterprise Content Management, Geschäftsprozess- und Kundenbeziehungsmanagement, Stammdatenmanagement und Enterprise Resource Planning, Cloud-Lösungen für unterschiedliche Unternehmensbereiche
- Quantencomputing, u. a. Anwendungspotenziale, Algorithmenentwicklung und -auswahl, Softwareengineering, Entwicklungswerkzeuge, Benchmarking und Hardwareevaluation, Vorgehensweisen, umfassendes Schulungsprogramm

Das Autorenteam

Harriet Kasper ist Senior Beraterin beim Fraunhofer IAO in Stuttgart und moderiert unter anderem Innovationsnetzwerke. Ihre Schwerpunkte sind Technologie-Foresight sowie Konzeption von Digitalisierungslösungen. Im Zentrum stehen dabei die Einsatzmöglichkeiten von Künstlicher Intelligenz, geeignete Interaktionsformate und Maßnahmen zur Akzeptanz von KI-Systemen.

Harriet Kasper
Fraunhofer IAO
Forschungsbereich
Digital Business
Team Angewandte KI
harriet.kasper@
iao.fraunhofer.de

Dennis Klau ist Experte für Machine Learning und Generative KI. Er ist im Rahmen des KI-Fortschrittszentrums an der Konzeption und Umsetzung mehrerer RAG-Systeme beteiligt und gibt sein Wissen in Schulungen und Seminaren weiter. Im Forschungs- und Innovationszentrum für Hybride Künstliche Intelligenz leitet er die technische Entwicklung.

Dennis Klau
Fraunhofer IAO
Forschungsbereich
Digital Business
Team Angewandte KI
dennis.klau@
iao.fraunhofer.de

KI-Fortschrittszentrum

Das KI-Fortschrittszentrum »Lernende Systeme und Kognitive Robotik« unterstützt Firmen dabei, die wirtschaftlichen Chancen der Künstlichen Intelligenz und insbesondere des Maschinellen Lernens für sich zu nutzen. In anwendungsnahen Forschungsprojekten und in direkter Kooperation mit Industrieunternehmen arbeiten die Stuttgarter Fraunhofer-Institute für Produktionstechnik und Automatisierung IPA sowie für Arbeitswirtschaft und Organisation IAO daran, Technologien aus der KI-Spitzenforschung in die breite Anwendung der produzierenden Industrie und der Dienstleistungswirtschaft zu bringen. Unterstützt werden sie dabei vom Institut für Arbeitswissenschaft und Technologiemanagement IAT der Universität Stuttgart. Finanzielle Förderung erhält das Zentrum vom Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg.

Mission
Das KI-Fortschrittszentrum ist der anwendungsorientierte Zweig von Cyber Valley, Europas größter Forschungskooperation im Bereich der Künstlichen Intelligenz. Darüber hinaus ist das KI-Fortschrittszentrum Bestandteil von S-TEC, dem Stuttgarter Technologie- und Innovationscampus: www.s-tec.de

Das KI-Fortschrittszentrum schlägt die Brücke von der KI-Spitzenforschung in den Mittelstand und macht KI-Technologien für die Wirtschaft in Baden-Württemberg und darüber hinaus nutzbar. Als führender Innovationspartner für den Mittelstand arbeitet das Zentrum an Themen, die für den Einsatz von KI und Robotik branchenübergreifend von zentraler Bedeutung sind, beispielsweise Autonomie, Effizienz und Nachhaltigkeit, Mensch-Maschine-Interaktion sowie Vertrauen. Das KI-Fortschrittszentrum informiert Unternehmen über Technologietrends und deren Einsatzpotenziale und unterstützt sie bedarfsgerecht und niedrigschwellig bei der Entwicklung und Umsetzung von ambitionierten KI-Innovationen, damit sie die wirtschaftlichen Chancen der KI künftig noch besser nutzen können.

Vision
Das KI-Fortschrittszentrum ist ein Leuchtturm für erfolgreichen Technologietransfer in den Mittelstand und ermöglicht Unternehmen einen wirtschaftlichen und verantwortungsvollen Einsatz von Künstlicher Intelligenz und Robotik für unternehmerischen Erfolg sowie individuellen und gesellschaftlichen Nutzen.

Studienreihe »Lernende Systeme und Kognitive Robotik«
Die Studienreihe »Lernende Systeme und Kognitive Robotik« gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Nähere Informationen und die aktuellen Versionen der Studien finden Sie unter: www.ki-fortschrittszentrum.de/de/themen/studien.html

Impressum

Kontaktadresse
Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO
Nobelstraße 12, 70569 Stuttgart

Harriet Kasper
Angewandte Künstliche Intelligenz
Telefon +49 711 970-2357
harriet.kasper@iao.fraunhofer.de

Dennis Klau
Angewandte Künstliche Intelligenz
Telefon +49 711 970-5112
dennis.klau@iao.fraunhofer.de

Fraunhofer-Publica
http://dx.doi.org/10.24406/publica-7451

Satz und Layout
Franz Schneider, Fraunhofer IAO

Titelbild
© Tameem – Adobe Stock/Fraunhofer IAO

Gefördert durch das Ministerium für Wirtschaft, Arbeit
und Tourismus Baden-Württemberg

Kontakt

KI-Fortschrittszentrum
Fraunhofer IAO und Fraunhofer IPA
Nobelstraße 12
70569 Stuttgart
www.ki-fortschrittszentrum.de

Harriet Kasper
Telefon +49 711 970-2357
harriet.kasper@iao.fraunhofer.de

Fraunhofer IAO
Nobelstraße 12
70569 Stuttgart
www.iao.fraunhofer.de

Gefördert durch

Baden-Württemberg
Ministerium für Wirtschaft,
Arbeit und Tourismus

Partner

Cyber Valley

S·TEC

Universität Stuttgart
Institut für Arbeitswissenschaft und
Technologiemanagement IAT