Ein praktischer Leitfaden zur Implementierung von Generativer KI für Retrieval-Augmented Generation

Atos
Research
Community

Inhaltsverzeichnis

Executive Summary

Laut Vorhersagen von Gartner sieht die Zukunft der generativen KI (GenAI) vielversprechend aus, mit einer starken Verbreitung in Unternehmen in den kommenden Jahren. Bis 2026 wird erwartet, dass über 80% der Unternehmen generative KI-Programmierschnittstellen (APIs) und Modelle nutzen oder generative KI-unterstützte Anwendungen in Produktionsumgebungen einsetzen werden. Dies ist ein signifikanter Anstieg gegenüber weniger als 5% heute[1].

Die Prognosen deuten darauf hin, dass die Abhängigkeit verschiedener Sektoren von generativer KI steigt, was die Arbeitsweise von Unternehmen und die Interaktion von Individuen mit Technologie revolutioniert.

In dieser sich wandelnden Umgebung tritt die Retrieval-Augmented Generation (RAG) als bewährter Ansatz für Anwendungen in Unternehmen hervor. RAG ist eine Technik, die Large Language Models (LLMs) verbessert, indem externe Wissensquellen integriert werden. Durch die zusätzlichen Informationen wird die Leistung des LLMs verbessert und die Antworten werden genauer und fundierter. Indem RAG Unternehmen in die Lage versetzt, eine Flut von Daten zu nutzen und in verwertbare Informationen umzuwandeln, stellt es eine gute Lösung für Organisationen dar, die mit Datenüberfluss und der Notwendigkeit präzisionsgetriebener Entscheidungsfindung zu kämpfen haben.

Dieses Whitepaper dient als Kompass, um das transformative Potenzial von RAG zu erschließen und einen Weg für dessen Integration in Geschäftsstrategien und -operationen aufzuzeigen. Es untersucht, wie das volle Potenzial von RAG und den damit verbundenen Lösungen dieser Spitzentechnologie genutzt werden kann, um Organisationen zu befähigen, in einer zunehmend datengetriebenen und KI-gestützten Welt immer einen Schritt voraus zu sein.

RAG: Eine Verschmelzung von Daten und Dialog

RAG repräsentiert im Kern eine Verschmelzung von zwei Bereichen der Künstlichen Intelligenz: Data Retrieval und Natural Language Generation. Es ermöglicht Maschinen, in die Tiefen digitaler Daten abzutauchen und mit neuen Erkenntnissen zurückzukehren, die durch ein Sophisticated Language Model ausgedrückt werden. In dieser Konfiguration ist RAG nicht nur ein Werkzeug, sondern ein Partner, der menschliche Expertise mit datengetriebenem Scharfsinn erweitert.

Strategische Imperative für Unternehmensleiter

Für strategisch denkende Geschäftsführer eröffnet RAG Wege zur verbesserten Kundeninteraktion, präziseren Markteinblicken und optimierte interne Kommunikation. Die Auswirkungen sind tiefgreifend. Marketingkampagnen können mit beispielloser Präzision auf die Zielgruppe zugeschnitten werden, der Kundenservice kann sich zu einem Dialog entwickeln, der durch den vollständigen Kontext der bisherigen Kundenreise unterstützt wird, und Führungskräfte können Entscheidungen mit dem Gewicht umfassender Datenanalysen treffen.

Überwindung von Implementierungshürden

Doch wie bei jeder neuen Technologie gibt es auch bei der Einführung von RAG viele Hürden: technische Herausforderungen, begrenzte Ressourcen und ethisch-moralische Fragen. Dieses Dokument beleuchtet nicht nur diese Herausforderungen, sondern präsentiert auch pragmatische Lösungen, von der Nutzung der Infrastruktur von Cloud-Plattformen bis hin zur Implementierung strenger ethischer Standards, die den Einsatz von RAG-Systemen leiten.

Handlungsleitfaden

Was folgt, ist ein umfassender Handlungsleitfaden für zukunftsorientierte Unternehmen, die bereit sind, RAG-Technologie einzusetzen. Im Folgenden werden die Grundprinzipien analysiert, die Anwendungen bewertet und die Entwicklung prognostiziert, alles aus einer pragmatischen, geschäftszentrierten Perspektive. Die unten enthaltenen Erkenntnisse basieren auf Branchenexpertise, Fallstudien und einer visionären Sichtweise auf die Rolle der KI bei der Gestaltung der Zukunft von Unternehmen.

[1]Gartner Experts Answer the Top Generative AI Questions for Your Enterprise; https://www.gartner.com/en/topics/generative-ai

Einführung in die Retrieval Augmented Generation

RAG ist eine Technik, die die Fähigkeiten von Large Language Modellen (LLMs) mit externen Wissensquellen kombiniert, um fundiertere und faktenbasierte Ausgaben zu erzeugen. Durch das Abrufen relevanter Informationen aus Datenbanken, Dokumenten oder dem Internet erweitert RAG das Wissen des LLMs und ermöglicht es, Antworten zu erzeugen, die aktuelle und domänenspezifische Informationen enthalten. Es gibt verschiedene Ansätze zur Implementierung von RAG, die jeweils ihre eigenen Stärken und geeigneten Anwendungsfälle haben.

Ansatz	Beschreibung	Geeignete Anwendung	Implementierungsbeispiel
Fine-Tuning	Weiterführung des Trainingsprozesses eines vortrainierten LLMs auf einem kleineren, spezialisierten Datensatz, um dessen Antworten auf spezifische Domänen oder Anwendungen abzustimmen.	Anwendungen, die hohe Genauigkeitsanforderungen in speziellen Bereichen erfordern, wie beispielsweise medizinische Diagnosen oder rechtliche Analysen.	Das Training eines LLMs anhand von medizinischen Aufzeichnungen zur Verbesserung seiner Leistung bei der Generierung genauer medizinischer Diagnosen.
Prompt-Tuning	Verwendung eines kleinen trainierbaren Modells zur Kodierung von Texteingaben und zur Generierung aufgabenspezifischer virtueller Tokens, die die Antworten des LLMs in Richtung des gewünschten Outputs lenken.	Aufgaben, die erfordern, dass ein Modell auf eine spezifische Weise agiert, ohne umfangreich neu trainiert zu werden, vor allem wenn es sich um eine Vielzahl von Aufgaben handelt, die durch Prompts gesteuert werden können.	Optimierung eines virtuellen Assistenten zur Handhabung verschiedener Gesprächsstile, wie formaler Kundensupport und zwanglose Benutzerinteraktionen, durch Anpassung der Prompts.
Low-Rank Adaptation (LoRA)	Anpassung eines vortrainierten Modells, um es besser auf einen spezifischen Datensatz abzustimmen, indem nur eine kleine Auswahl der Schlüsselparameter modifiziert wird, wodurch effiziente und kostengünstige Anpassungen ermöglicht werden.	Aufgaben, bei denen es wenige domänenspezifischen Daten gibt und die Rechenressourcen begrenzt sind, bewahrt es die allgemeinen Fähigkeiten des LLM, während es gleichzeitig auf spezifische Arten von Eingaben flexibler reagiert.	Anpassung eines allgemein einsetzbaren LLMs für die Analyse rechtlicher Dokumente durch Training auf juristischer Fachsprache und Konzepte, ohne dabei umfassende sprachliche Fähigkeiten zu verlieren.
Retrieval-Augmented Generation (RAG)	Die Kombination der Fähigkeiten von LLMs mit externen Wissensquellen, um relevante Informationen in Echtzeit abzurufen und diese zur Erweiterung der Antworten des Modells zu nutzen.	Anwendungen, bei denen das LLM aktuelle Informationen bereitstellen muss oder wenn die Aufgabe Wissen erfordert, das nicht in den Trainingsdaten des LLM enthalten ist.	Entwickeln von Tools, die Zusammenfassungen der neuesten Nachrichtenartikel bereitstellen können, indem sie aktuelle Nachrichtenfeeds abrufen und in ihre Ausgabe integrieren.

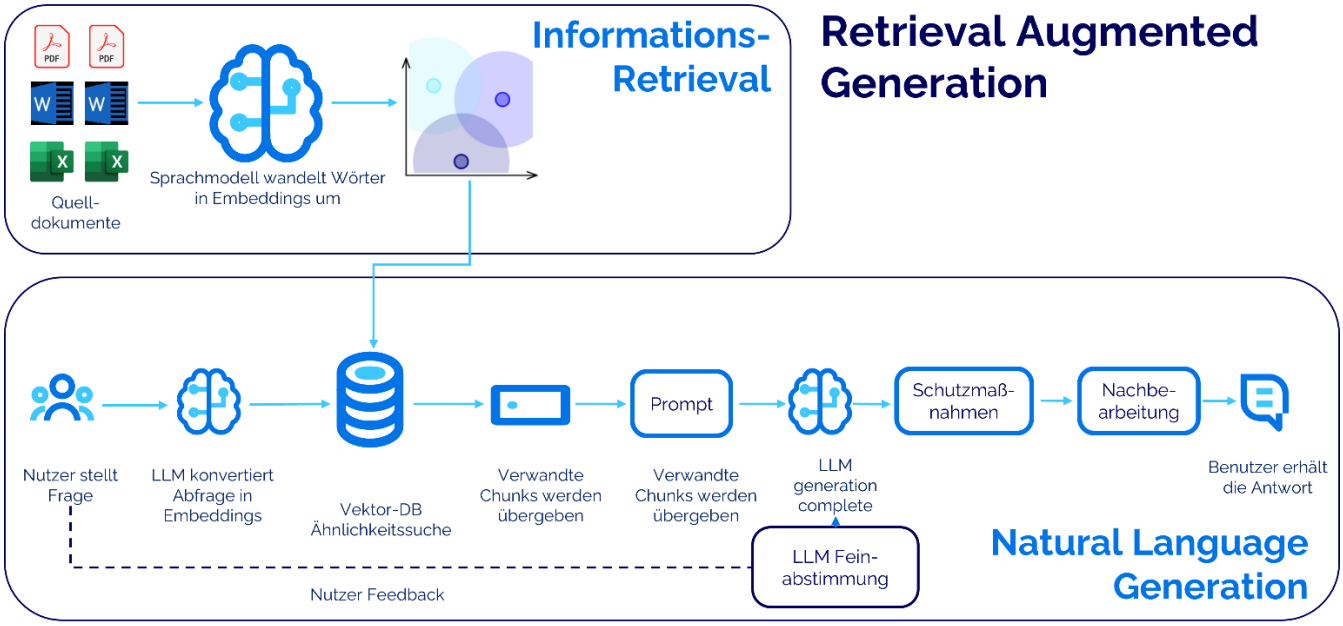
Die Wahl des Ansatzes hängt von Faktoren wie der Notwendigkeit der Domänenspezifität, der Recheneffizienz, der Notwendigkeit aktueller Informationen und dem Umfang der erforderlichen Anpassung ab. Im nächsten Kapitel wird der RAG-Ansatz vertieft und dabei die Kernkomponenten, Implementierungsstrategien und praktischen Anwendungen untersucht.

Die Technologie hinter RAG: Ein tiefgehender Einblick

Enthüllung der Mechanismen von RAG-Systemen

Retrieval-Augmented Generation (RAG) operiert an der Schnittstelle von zwei fortschrittlichen Technologien: Information Retrieval (IR) und Natural Language Generation (NLG). Um die Innovation, die RAG mit sich bringt, vollständig zu erfassen, muss man das komplexe Zusammenspiel dieser beiden Komponenten verstehen.

Im Folgenden wird die High-Level-Architektur eines typischen RAG-Systems dargestellt:



Information Retrieval: Die Suche nach Relevanz

Die IR-Komponente von RAG ist vergleichbar mit einem Bibliothekar, der über enzyklopädisches Wissen und sofortigen Zugriff darauf verfügt. Sie durchforstet Terabytes an Daten – sowohl strukturierte als auch unstrukturierte – um die relevantesten Informationen als Antwort auf eine Anfrage zu finden. Moderne IR-Systeme basieren auf fortschrittlichen Algorithmen, einschließlich maschineller Lernmodelle, die darauf trainiert wurden, die Semantik einer Suchanfrage zu verstehen, nicht nur deren Schlüsselwörter.

In geschäftlichen Anwendungen müssen IR-Systeme besonders geschickt darin sein, branchenspezifische Fachsprache und Nuancen zu verstehen. Sie müssen auch in der Lage sein, die Absicht hinter Anfragen zu erkennen, die nicht immer explizit formuliert sind. Dieses Verständnis ermöglicht es Unternehmen, nicht nur genaue Informationen abzurufen, sondern auch Erkenntnisse zu gewinnen, die wirklich relevant für die jeweilige Anfrage sind.

Natural Language Generation: Die Kunst der Artikulation

Sobald die relevanten Informationen abgerufen wurden, übernimmt die NLG-Komponente. Sie ist dafür verantwortlich, kohärente, fließende und zum Kontext passende Antworten aus den abgerufenen Daten zu erstellen. Angetrieben durch Fortschritte im Deep Learning – insbesondere mit Transformer basierten Modellen wie GPT (Generative Pre-trained Transformer) – hat NLG neue Höhen in ihrer Fähigkeit erreicht, menschenähnlichen Text zu generieren.

In einem RAG-System generiert die NLG-Komponente den Text nicht einfach im Vakuum; sie setzt die Informationen in Kontext der umfassenderen Anfrage des Nutzers. Für Unternehmen bedeutet dies, Inhalte zu produzieren, die nicht nur genau und informativ sind, sondern auch ansprechend und im Einklang mit der Unternehmenssprache und den Kommunikationsstandards stehen.

Die Synergie zwischen IR und NLG

Die wahre Stärke von RAG liegt im Zusammenspiel zwischen IR und NLG. Diese Synergie wird durch eine dynamische Feedback-Schleife erreicht, bei der die NLG-Komponente den Suchprozess der IR-Komponente beeinflussen kann und umgekehrt. So kann die NLG-Komponente eine vorläufige Antwort generieren, die das IR-System verwendet, um seine nachfolgenden Suchvorgänge zu verfeinern, wodurch die Genauigkeit und Relevanz der abgerufenen Informationen verbessert wird.

Im Kontext von Business-Intelligence bedeutet diese synergetische Schleife, dass RAG-Systeme sich an wandelnde Datenlandschaften anpassen können, ihre Ausgaben in Echtzeit verfeinern, sobald neue Informationen verfügbar werden oder sich der Kontext des Gesprächs ändert.

Überwindung technischer Herausforderungen

Trotz der Raffinesse der RAG-Technologie ist sie nicht ohne Herausforderungen. Die Integration externer Datenquellen erfordert eine sorgfältige Abstimmung der Informationen. Das System muss zwischen relevanten und überflüssigen Daten unterscheiden, um sicherzustellen, dass die generierten Antworten Kohärenz und Kontext beibehalten. Darüber hinaus erfordert die Komplexität dieser Systeme ein empfindliches Gleichgewicht zwischen Recheneffizienz und dem Detailreichtum der Ausgaben.

Die technischen Erfolge von RAG im Geschäftsumfeld

Wenn diese Herausforderungen erfolgreich gemeistert werden, können RAG-Systeme Geschäftsabläufe transformieren. Im Kundenservice können RAG-Systeme als Komponenten in Agentensysteme integriert werden, um präzise Informationen bereitzustellen, wodurch die Lösungszeiten verkürzt und die Kundenzufriedenheit erhöht werden. In der Marktanalyse können sie große Datenmengen schnell synthetisieren und Erkenntnisse liefern, die Analysten Tage oder Wochen kosten könnten, um sie zu entdecken.

Die Anwendung von RAG im Geschäftsbereich

RAG: Ein vielseitiges Werkzeug für verschiedene Branchen

Die Vielseitigkeit von Retrieval-Augmented Generation (RAG) liegt in ihrer Fähigkeit, sich an eine Vielzahl von Geschäftsbereichen anzupassen und diese zu erweitern. Hier werden die transformative Wirkung von RAG in verschiedenen Bereichen untersucht, die jeweils einzigartige Herausforderungen und Chancen aufweisen.

Verbesserung der Kundenerfahrung

Im Bereich des Kundenservice kann RAG ein echter Game-Changer sein. Fortschrittliche Chatbots und Support-Systeme, die von RAG betrieben werden, können Kunden schnelle, genaue und kontextrelevante Informationen bereitstellen. Beispielsweise kann ein RAG-gestütztes System in einem Telekommunikationsunternehmen kundenspezifische Daten wie Nutzungsmuster und Abrechnungshistorie abrufen und personalisierte Antworten generieren, die individuelle Anliegen präzise adressieren.

Optimierung von Forschung und Entwicklung

F&E-Teams können enorm von RAGs Fähigkeit profitieren, große Mengen an Informationen aus Forschungspapieren, Patenten und technischen Dokumenten zu synthetisieren und zusammenzufassen. Es ermöglicht Forschern, über die neuesten Entwicklungen auf dem Laufenden zu bleiben, ohne sich durch zahllose Artikel arbeiten zu müssen, wodurch Innovationen beschleunigt und die Markteinführungszeit für neue Produkte verkürzt werden.

Optimierung von Content-Erstellung und Marketing

Marketingexperten, die RAG nutzen, können Inhalte erstellen, die bei ihrer Zielgruppe tiefen Anklang finden. Durch das Abrufen und Generieren von Erkenntnissen aus Kundendaten, Markttrends und Wettbewerbsanalysen kann RAG dabei helfen, hochgradig zielgerichtete Kampagnen zu erstellen, die direkt auf die Bedürfnisse und Wünsche der Kunden eingehen.

Navigieren in komplexen rechtlichen Landschaften

Juristen können RAG nutzen, um Rechtsprechung und Gesetze zu durchsuchen, Zusammenfassungen zu erstellen und relevante Präzedenzfälle für laufende Fälle zu identifizieren. Dies verbessert die Effizienz der juristischen Recherche und unterstützt fundierte Entscheidungen in komplexen rechtlichen Landschaften.

Finanzanalyse und Reporting

Im Finanzwesen können RAG-Systeme Marktdaten, Finanzberichte und Wirtschaftsindikatoren analysieren und dabei prägnante, aufschlussreiche Zusammenfassungen liefern. Diese Anwendung ist entscheidend für rechtzeitige Investitionsentscheidungen und die Formulierung strategischer Finanzpläne.

Individuelle Anwendungen in Nischenbereichen

Die Anpassungsfähigkeit von RAG zeigt sich in seinen individuellen Anwendungen in Nischenbereichen. Im Gesundheitswesen kann RAG beispielsweise Patientenaufklärungsmaterialien erstellen, indem Informationen aus einer umfangreichen medizinischen Wissensdatenbank abgerufen und auf individuelle Bedürfnisse und das Verständnisniveau der Patienten zugeschnitten werden.

Herausforderungen bei vielfältigen Anwendungen

Die Implementierung von RAG-Systemen in diesen vielfältigen Anwendungen ist nicht ohne Herausforderungen. Jedes Gebiet hat seine eigenen Besonderheiten bei Daten, Datenschutzbedenken und Genauigkeitsanforderungen. Der Schlüssel zu einer erfolgreichen Implementierung liegt in einem tiefen Verständnis der Nuancen des jeweiligen Gebiets und einer strengen Qualitätskontrolle, um sicherzustellen, dass das RAG-System die hohen Standards erfüllt, die von Fachleuten in diesem Bereich erwartet werden.

Geschäftstransformation durch RAG

Insgesamt fungiert RAG als Säule der digitalen Transformation für Unternehmen und treibt sie in neue Dimensionen der Effizienz und Personalisierung. Durch die Integration von RAG in ihre Abläufe können Unternehmen nicht nur ihre Dienstleistungsangebote verbessern, sondern auch neue Wertschöpfungspotenziale erschließen, die zuvor unerreichbar waren.

Ein genauerer Blick: Potenzielle Anwendungen für RAG

Unternehmenswissen Q&A

RAG kann dabei helfen, relevante Informationen aus Wissensdatenbanken abzurufen und dadurch kontextbezogene und genaue Antworten auf Anfragen zu ermöglichen.

Automatisierte Angebotsantworten

RAG kann Inhalte für spezifische Angebotsantworten generieren, die aus Texten, Bildern und Videos in Ihrer Unternehmensdatenbank stammen.

Unternehmensspezifische Chatbots

Eine RAG-basierte Lösung kann Chatbots befähigen, unternehmensspezifische Informationen aus Ihrer Wissensdatenbank oder anderen relevanten Quellen abzurufen.

Produktmanagement

Durch das Zusammenführen von Kundenfeedback, Markttrends und Produktleistungsdaten kann RAG Produktmanager in die Lage versetzen, datengestützte Entscheidungen für Produktverbesserungen und Innovationsinitiativen zu treffen.

Alternative Kanalsuchen (wie MS Teams)

Extrahieren von Informationen und Generieren von Metadaten aus Dokumenten, die in Kanälen wie Microsoft Teams geteilt werden. Diese Metadaten können Tags, Zusammenfassungen, Schlüsselkonzepte oder sogar vorgeschlagene verwandte Dokumente umfassen.

Herausforderungen und strategische Lösungen bei der RAG-Implementierung

Die Implementierung von Retrieval-Augmented Generation (RAG) innerhalb einer Organisation bringt eine Reihe von Herausforderungen mit sich, die technische, ethische und sicherheitsrelevante Aspekte umfassen. Diese Herausforderungen zu verstehen und anzugehen, ist entscheidend, um das volle Potenzial von RAG auszuschöpfen.

Technische Feinheiten: Daten- und Modelloptimierung

Eine der wichtigsten Überlegungen bei der Implementierung von RAG ist die Integration externer Datenquellen mit KI-Modellen. Die nahtlose Verbindung von Datenabruf und Textgenerierung erfordert:

- Datenqualitätsmanagement (DQM): Sicherstellung der Genauigkeit und Aktualität der Informationen in der Datenbank, was kontinuierliche Aktualisierungen und Monitoring erfordert.
- Modelltraining und -tuning: Entwicklung eines Modells, das nicht nur die Nuancen der Daten versteht, sondern auch kohärente, kontextbewusste Antworten generiert.

Security- und Compliance-Maßnahmen

Die Integration von RAG in Sektoren mit strengen Security- und Compliance-Anforderungen umfasst:

- Datenschutzmechanismen: Verschlüsselung sensibler Daten und Nutzung sicherer Kommunikationskanäle.
- Einhaltung von Vorschriften: Sicherstellung, dass das RAG-System mit branchenspezifischen Vorschriften und Standards wie der DSGVO für den Datenschutz und dem EU AI Act übereinstimmt.

Ressourcenzuweisung und Skalierungsstrategien

RAG-Systeme sind ressourcenintensiv und erfordern erhebliche Rechenleistung und Speicherplatz. Strategische Lösungen umfassen:

- Cloud-Computing: Nutzung von Cloud-Ressourcen für eine skalierbare Infrastruktur, die sich an die schwankenden Anforderungen der RAG-Operationen anpasst.
- Effiziente Algorithmen: Implementierung von Algorithmen, die den Nutzen der Rechenressourcen maximieren, ohne die Leistung zu beeinträchtigen.

Ethische Überlegungen: Verbesserung von Bias und Fairness

RAG-Systeme sind, wie alle KI-Technologien, anfällig für Vorurteile, die in ihren Trainingsdaten vorhanden sind. Um dies anzugehen, sind folgende Maßnahmen erforderlich:

- Vielfältige Datensätze: Aufbau von Trainingssätzen aus einer breiten Palette von Quellen, um das Risiko von Vorurteilen zu minimieren.
- Bias-Erkennungsalgorithmen: Einsatz spezialisierter Algorithmen zur Erkennung und Mitigierung von Vorurteilen im System.

Datenschutzbedenken und verantwortungsvolle KI-Praktiken

Der Umgang mit sensiblen Informationen und die verantwortungsvolle Implementierung von KI erfordern:

- Datenschutzwahrende Techniken: Implementierung von Methoden wie Datenanonymisierung und sicherer Datenspeicherung, um die Privatsphäre der Nutzer zu schützen.
- Ethische Richtlinien: Etablierung klarer Richtlinien für den verantwortungsvollen Einsatz von KI, basierend auf den neuesten Forschungsergebnissen und ethischen Standards.

Überwachbarkeit und Leistungsmonitoring

Die Aufrechterhaltung einer hohen Leistung und die Gewährleistung der Sicherheit von RAG-Systemen erfordern Überwachbarkeit – das Monitoring der Systemgesundheit, der Benutzerinteraktionen und der Datenflüsse. Zu den erforderlichen Lösungen gehören:

- Erweiterte Analytik: Einsatz von Analysetools zum Real-Time-Monitoring der Systemleistung und der Benutzerinteraktionen.
- Automatisierte Alarmsysteme: Einrichtung automatisierter Alarme bei Anomalien, um eine schnelle Reaktion auf potenzielle Probleme zu gewährleisten.

Ein strategischer Ansatz zur Implementierung von RAG umfasst:

- Funktionsübergreifende Teams: Zusammenstellung von Teams mit unterschiedlicher Expertise, um die Integration von RAG-Systemen zu überwachen und sicherzustellen, dass technische, ethische und sicherheitsrelevante Aspekte gleichermaßen berücksichtigt werden.
- Pilotprogramme und schrittweise Einführung: Testen von RAG-Systemen in kleinerem Maßstab vor der breiteren Einführung, um potenzielle Herausforderungen zu identifizieren und die Leistung zu optimieren.

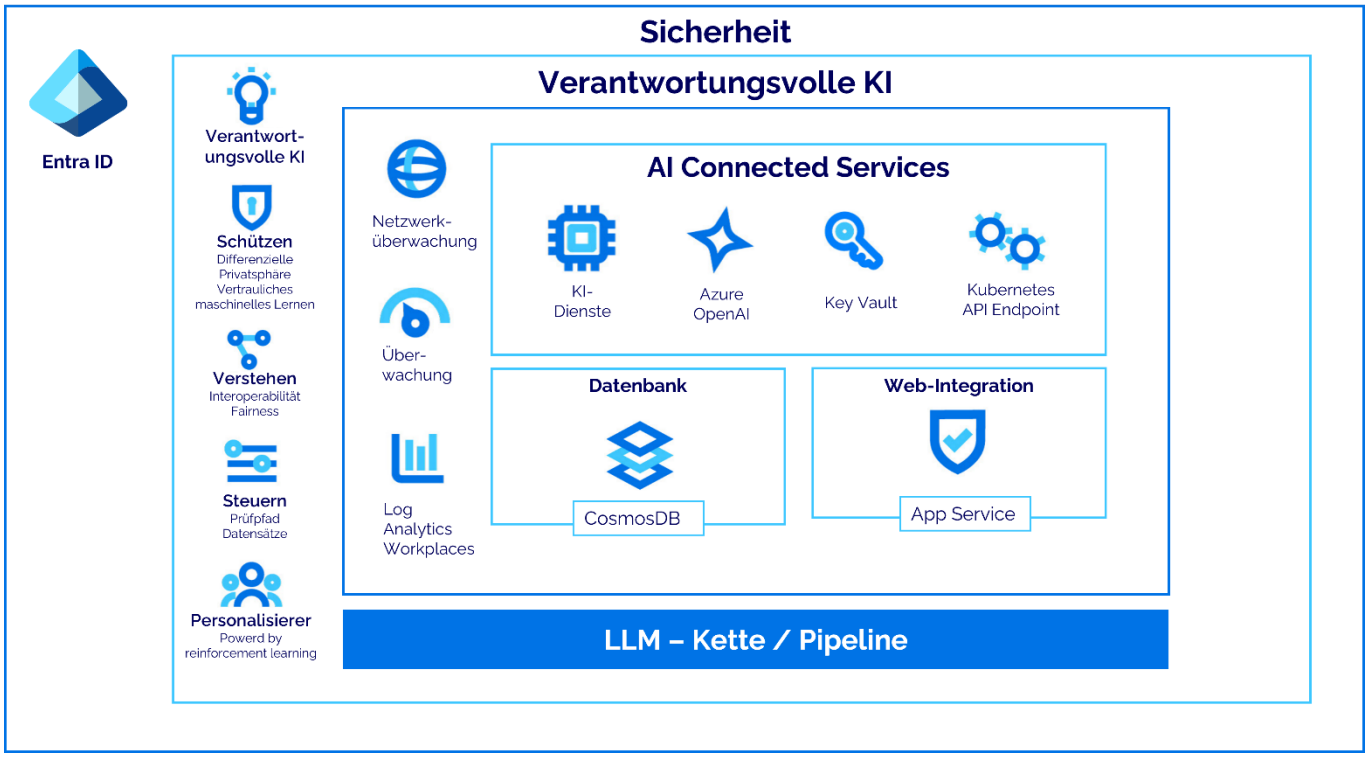
Überwindung der Herausforderungen bei der RAG-Implementierung: Eine Fallstudie

Um die zahlreichen Herausforderungen im Zusammenhang mit RAG-Systemen besser zu verstehen, wurde eine Plattform entwickelt, die mehrere LLMs orchestriert und diese Plattform für verschiedene Kunden implementiert. Unsere praktische Erfahrung war entscheidend für die Entwicklung und Implementierung einer effektiven RAG-Lösung. Es wurden zahlreiche Plattformen evaluiert und aufgrund ihrer integrierten Suite von Diensten, die die RAG-Bereitstellung erleichtern, für Microsoft Azure[2] entschieden.

Es ist keineswegs die einzige Plattform, die eine RAG-Bereitstellung unterstützt, sondern lediglich diejenige, die für die Erstellung dieses RAG-Testumfelds als am besten geeignet erachtet wurde. Eine detaillierte Evaluierung aller potenziellen Plattformen liegt außerhalb des Umfangs dieses Papiers. Andere alternative Plattformen umfassen AWS Bedrock, Google Vertex oder Open-Source-Plattformen, und ermutigt jedes Unternehmen, das RAG in Betracht zieht, die Plattform auszuwählen, die am besten zu Ihrer spezifischen Situation passt.

Im Folgenden zeigen wir, wie ein Unternehmen die technischen, ethischen und compliance-bezogenen Hürden von RAG meistern kann. In diesem Fall wurden die verfügbaren Funktionen und Dienste in Azure auf die im vorherigen Abschnitt beschriebenen Herausforderungen abgebildet.

Nachstehend sehen Sie eine High-Level-Architektur, die darstellt, wie eine verantwortungsbewusste LLMOps Azure RAG-Lösung implementiert wurde.



[2] Cloud Computing Services | Microsoft Azure

Integration und Skalierbarkeit

Azure bietet eine Vielzahl von Diensten, die die Integration von RAG-Komponenten ermöglichen:

- Azure AI Search: Treibt die Retrieval-Komponente an und ermöglicht das Katalogisieren großer Datenmengen für effiziente Abfragen.
- Azure Machine Learning: Ermöglicht die Erstellung, Bereitstellung und Wartung von Machine-Learning-Modellen im großen Maßstab.
- Azure Kubernetes Service (AKS): Bietet eine verwaltete Kubernetes-Umgebung für die Bereitstellung und Skalierung containerisierter RAG-Anwendungen.

Erweiterte Überwachungs- und Monitoring-Tools

Die Monitoring-Tools von Azure gewährleisten, dass jeder Aspekt des RAG-Systems transparent und beobachtbar ist:

- Azure Monitor and Application Insights: Bietet die Möglichkeit, die Anwendungsleistung und Benutzeraktivitäten zu überwachen, um sicherzustellen, dass Abweichungen von der erwarteten Leistung erkannt und behoben werden.
- Azure Log Analytics: Unterstützt beim Erfassen und Analysieren von Systemprotokollen und gewährt Einblicke in die Systemgesundheit von RAG-Bereitstellungen.

Ethische Rahmenwerke und Bias-minimierung

Azures Responsible AI Toolkit bietet Funktionen, die bei der Erstellung ethischer RAG-Systeme helfen:

- Azure AI Fairness Toolkit[3]: Unterstützt bei der Identifizierung und Minderung von Vorurteilen in KI-Modellen und stellt Fairness in automatisierten Entscheidungsprozessen sicher.
- Azure Personalizer: Bietet einen Reinforcement-Learning-Dienst, der personalisierte Benutzererfahrungen liefert und gleichzeitig ethische Überlegungen berücksichtigt.

Datenschutz und Sicherheit: Zentrale Prioritäten

Azure bietet integrierte Datenschutz-, Sicherheits- und Compliance-Funktionen:

- Azure Security Center: Bietet einheitliches Sicherheitsmanagement und erweiterten Bedrohungsschutz für hybride Cloud-Workloads.
- Entra ID[4] — ehemals Azure Active Directory (AAD): Erzwingt starke Authentifizierung und rollenbasierte Zugriffskontrolle und schützt so den Zugang zu RAG-Systemen.
- Compliance: Azure hält eine umfangreiche Liste von Compliance-Zertifizierungen[5] aufrecht, die es ermöglichen, dass RAG-Bereitstellungen globale regulatorische Standards wie GDPR und HIPAA erfüllen.

Verwirklichung des Potenzials von RAG durch den Azure OpenAI Service

Der Azure OpenAI Service bietet Zugriff auf KI-Modelle, einschließlich der von OpenAI entwickelten, und integriert die Sicherheits- und Compliance-Ebenen von Azure. Dies ermöglicht den Aufbau von RAG-Systemen, die sowohl intelligent als auch im Einklang mit den Governance- und Richtlinienanforderungen der Organisation stehen.

Das Ergebnis: Eine synergetische RAG-Implementierung

Unser Proof of Concept zeigte, dass Azure-Dienste genutzt werden können, um ein robustes und vielseitiges RAG-System zu entwickeln, das in der Lage ist, mit den wachsenden Geschäftsanforderungen zu skalieren. Unabhängig davon, welche Cloud-Plattform gewählt wird, müssen Organisationen sicherstellen, dass ihre RAG-Systeme auf einer Grundlage von Sicherheit, Compliance und ethischer Integrität aufgebaut sind. Dies ebnet den Weg für innovative Anwendungen, die die Grenzen dessen erweitern, was KI im Geschäftsbereich leisten kann.

---

[3] Assess ML models' fairness in Python (preview) - Azure Machine Learning | Microsoft Learn
[4] Microsoft Entra ID (formerly Azure Active Directory) | Microsoft Security
[5] Azure compliance documentation | Microsoft Learn

Strategische Implementierung von RAG: Best Practices und Integration von MLOps

Die erfolgreiche Implementierung von Retrieval-Augmented Generation (RAG) innerhalb einer Organisation hängt von einer gut durchdachten Strategie ab, die Datenmanagement, Modellkonfiguration und kontinuierliches Monitoring umfasst.

Machine Learning Operations (MLOps) dient als Grundlage, und mit dem Aufkommen generativer KI – insbesondere Retrieval-Augmented Generation – sind zusätzliche Funktionen innerhalb der MLOps-Pipeline erforderlich. Im Folgenden werden Best Practices und die Rolle von MLOps bei der effektiven Bereitstellung und Wartung von RAG-Systemen skizziert.

Die Weiterentwicklung von MLOps zur Unterstützung von RAG

- kennzeichnet vorhandene MLOps-Funktionen
+ kennzeichnet neue MLOps-Funktionen, die zur Unterstützung von RAG erforderlich sind

Foundation Model Management
- Vortraining und Ausrichtungsprozesse
- Modellversionierung und -registrierung
- Auswahl und Integration von Foundation model
- Modellkarten und Dokumentation
+ Prompt-Management

Daten- und Wissensmanagement
- Datenversionierung und Nachverfolgung der Herkunft
- Datenqualitätsüberwachung und Validierung
+ Generierung und Verwaltung synthetischer Daten
+ Einbettungsmanagement und Vektordatenbanken
+ RAG-spezifische Dokumentenverarbeitung und Chunking
+ Kontinuierliche Aufnahme neuer Datenquellen

Modellanpassung und -entwicklung
- Experimentverfolgung und Versionskontrolle
- Hyperparameter-Optimierung
+ Prompt-Engineering und -Management
+ Feinabstimmung auf Unternehmensdatensätzen
+ Reinforcement Learning from Human Feedback (RLHF) Pipelines

Bereitstellungs- und Inferenzpipeline
- CI/CD für Modelle und Anwendungen
- A/B-Tests und Canary Releases
+ Agent- und Chain-management
+ Implementierung von RAG-Workflows
+ Absicherung der Eingabe-/Ausgabeverarbeitung
+ Skalierbarkeit und Ressourcenmanagement

Monitoring and Performance Benchmarking
- Echtzeit-Leistungs Monitoring
- Erkennung von Daten- und Konzeptdrift
+ LLM-spezifische Metriken
+ RAG-spezifisches Monitoring (Retrieval-Qualität, Relevanz)
+ Monitoring von Chains- und Ausführung von Agenten
+ Benchmarking gegenüber Industriestandards

Governance, Ethik und Compliance
- Modelldokumentation und Erklärbarkeit
- Bias Erkennung und Fairness Metriken
- KI-Risikoanalyse und Mitigationstrategie
- Ethische KI-Richtlinien und deren Einhaltung
- • Reproduzierbarkeit und Prüfbarkeit von KI-Systemen

1. Foundation Model Management

Das Management von Grundlagenmodellen umfasst mehrere kritische Prozesse, beginnend mit Vortraining und Ausrichtungsprozessen, um sicherzustellen, dass die Modelle für spezifische Aufgaben angemessen konfiguriert und abgestimmt sind. Dazu gehört auch die Verwaltung von Modellversionen und Registern, um verschiedene Versionen und Aktualisierungen nachzuverfolgen. Darüber hinaus ist die Auswahl und Integration von Grundlagenmodellen, die für verschiedene Anwendungen geeignet sind, von entscheidender Bedeutung. Dokumentation, einschließlich Modellkarten, bietet detaillierte Einblicke in die Fähigkeiten und Einschränkungen der Modelle.

Ein weiterer Aspekt ist das Prompt-Management, das sich auf die Erstellung und Verwaltung von Prompts konzentriert, um die Ausgabe des Modells effektiv zu steuern. Effizientes Datenmanagement ist ebenfalls entscheidend und umfasst strukturierte Ansätze zur Sammlung, Speicherung und Katalogisierung der vom RAG-System verwendeten Daten. Regelmäßige Daten-Audits und Aktualisierungen stellen sicher, dass die Informationen relevant und genau bleiben.

2. Daten- und Wissensmanagement

Das Daten- und Wissensmanagement stellt die Integrität und Qualität der Daten durch Datenversionierung und Nachverfolgung der Herkunft sowie durch Datenqualitätsüberwachung und -validierung sicher. Dieser Bereich umfasst auch die Generierung und Verwaltung synthetischer Daten zur Erweiterung von Datensätzen. Die Verwaltung von Einbettungen und Vektordatenbanken ist entscheidend für einen effizienten Datenabruf.

RAG-spezifische Dokumentenverarbeitung und Chunking beinhalten das Aufteilen von Dokumenten, um sie besser für Retrieval-Augmented Generation-Modelle verarbeiten zu können. Die kontinuierliche Aufnahme neuer Datenquellen ist ebenfalls entscheidend, um die Wissensbasis aktuell und relevant zu halten. Darüber hinaus erweitert die Generierung und Verwaltung synthetischer Daten das Datenmanagement mit neuen generativen KI-Fähigkeiten, indem synthetische Trainingsdaten und Randfälle erstellt werden, um die Genauigkeit und Robustheit des Modells zu bewerten und zu zertifizieren. Das Ergänzen des Feature Stores mit Embedding-Informationen aus Vektordatenbanken umfasst die Darstellung von Datenproben als dichte, mehrdimensionale Vektoren und die Verwaltung dieser Embeddings in einer Vektordatenbank.

3. Modellanpassung und -entwicklung

Die Modellanpassung und -entwicklung umfasst das Verfolgen von Experimenten und die Verwaltung der Versionskontrolle, um die Reproduzierbarkeit sicherzustellen. Die Hyperparameter-Optimierung ist entscheidend für die Verbesserung der Modellleistung.

Das Prompt-Engineering und -Management konzentriert sich auf die Gestaltung und Verfeinerung von Prompts, um das Modellverhalten zu steuern. Das Fein-Tuning auf Unternehmensdatensätzen passt Modelle an spezifische organisatorische Bedürfnisse an. Reinforcement Learning from Human Feedback (RLHF)-Pipelines sind ebenfalls enthalten und nutzen menschliches Feedback zur iterativen Verbesserung der Modellleistung. Die Anpassung des RAG-Modells an geschäftliche Anforderungen umfasst die Auswahl geeigneter Algorithmen, das Training der Modelle mit domänenspezifischen Daten und die Feinabstimmung von Parametern für optimale Leistung.

4. Bereitstellungs- und Inferenzpipeline

Die Bereitstellungs- und Inferenzpipeline umfasst Prozesse für Continuous Integration/Continuous Deployment (CI/CD) für Modelle und Anwendungen, um nahtlose Updates und Integration sicherzustellen. A/B-Tests und Canary Releases helfen dabei, die Modellleistung vor der vollständigen Bereitstellung zu validieren. Die Verwaltung von Agents und Chains ist notwendig, um komplexe KI-Workflows zu koordinieren.

Die Implementierung von RAG-Workflows integriert Retrieval-Augmented Generation-Prozesse. Richtlinien für die Eingabe-/Ausgabeverarbeitung gewährleisten die Zuverlässigkeit und Sicherheit der Modellausgaben. Skalierbarkeit und Ressourcenmanagement sind ebenfalls entscheidend für die effiziente Handhabung variierender Lasten. Das Management von Agents und Chains definiert komplexe mehrstufige Anwendungslogik, kombiniert mehrere Grundlagenmodelle und APIs und erweitert Modelle mit externem Speicher und Wissen. Debugging, Tests und die Visualisierung des Ausführungsflusses sind entscheidend für ein effektives Management während des gesamten Lebenszyklus generativer KI.

5. Monitoring and Performance Benchmarking

Monitoring und Performance-Benchmarking umfasst die Echtzeit-Leistungsüberwachung, um sicherzustellen, dass Modelle effektiv arbeiten. Die Erkennung von Daten- und Konzeptdrift hilft, die Genauigkeit der Modelle über die Zeit hinweg zu erhalten. Spezifische Metriken für Large Language Models (LLMs) bieten Einblicke in deren Leistung.

RAG-spezifisches Monitoring bewertet die Qualität und Relevanz der abgerufenen Informationen. Das Chain and Agent Monitoring stellt den reibungslosen Betrieb komplexer KI-Systeme sicher. Das Benchmarking gegenüber Industriestandards hilft, wettbewerbsfähige Leistungsniveaus aufrechtzuerhalten. Kontinuierliches Monitoring ist entscheidend, wobei Key-Performance-Indicators (KPIs) und Analysetools Bereiche zur Verbesserung identifizieren. Strenge Testprotokolle, einschließlich Stresstests und Benutzerakzeptanztests, validieren die Leistung und stellen sicher, dass die Benutzeranforderungen erfüllt werden.

6. Governance, Ethik und Compliance

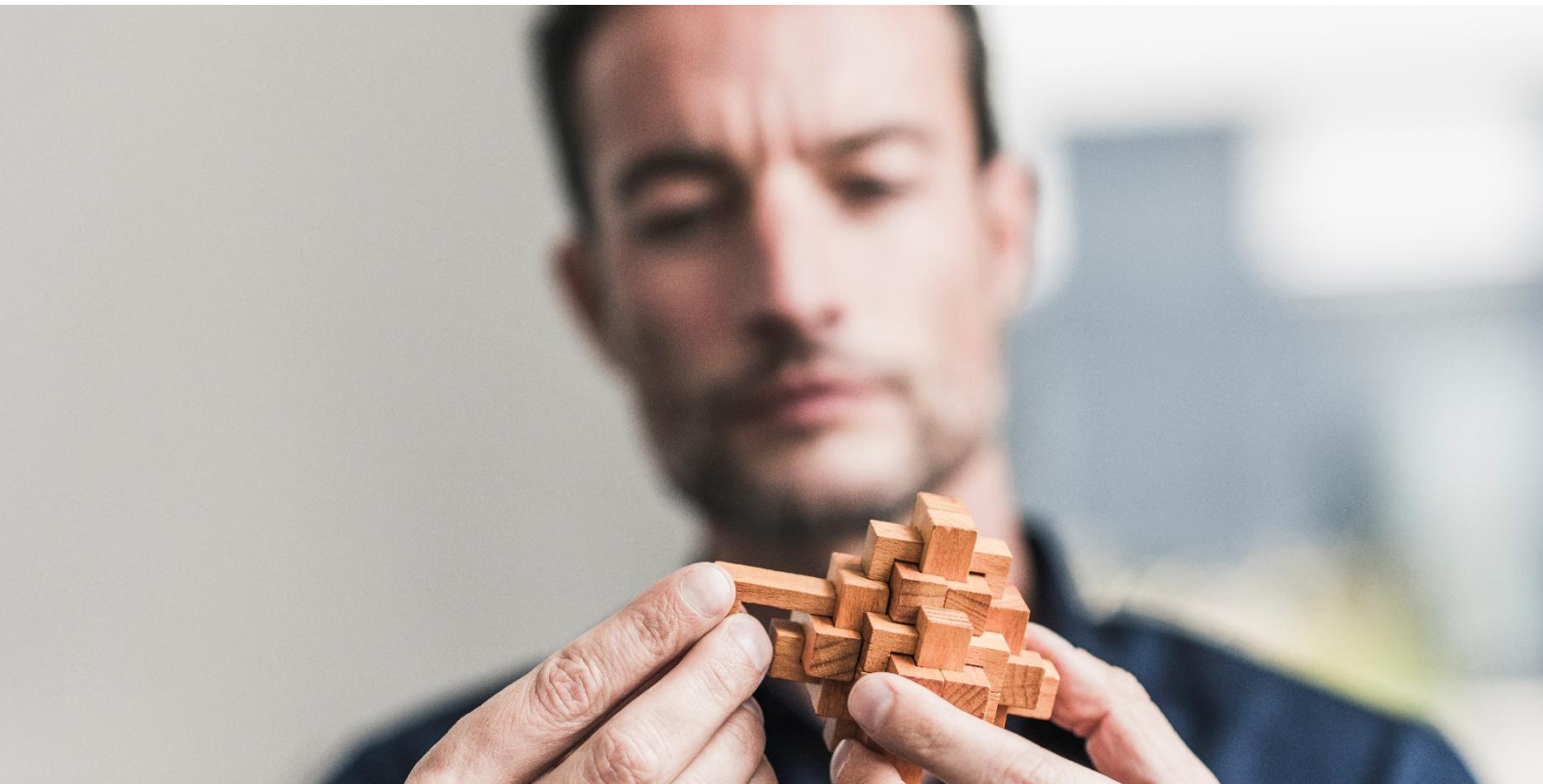
Governance, Ethik und Compliance stellen sicher, dass KI-Systeme verantwortungsvoll eingesetzt werden. Dazu gehören Modelldokumentation und Explainability, um Transparenz zu gewährleisten. Die Erkennung von Bias und Fairness Metriken sind entscheidend für die Einhaltung ethischer Standards. Die Bewertung und Mitigierung von KI-Risiken ist notwendig für eine sichere Implementierung.

Ethische KI-Richtlinien und deren Einhaltung umfassen das Befolgen von Best Practices und Vorschriften. Reproduzierbarkeit und Prüfbarkeit von KI-Systemen stellen sicher, dass KI-Prozesse zuverlässig repliziert und überprüft werden können. Die Integration von Governance und ethischen Überlegungen in die MLOps-Pipeline gewährleistet die Übereinstimmung mit den organisatorischen Werten und gesetzlichen Anforderungen. Kontinuierliches Monitoring auf Bias und die Implementierung von Korrekturmechanismen bewahren die ethische Integrität von RAG-Systemen.

MLOps: Orchestrierung der RAG-Implementierung

MLOps erleichtert die Erstellung automatisierter Workflows für das Training, die Validierung und die Bereitstellung von RAG-Modellen, wodurch ein nahtloser Übergang von Entwicklungs- zu Produktionsumgebungen gewährleistet wird. Skalierbarkeits- und Ressourcenmanagement-Tools helfen dabei, Rechenressourcen effizient zu verwalten. Versionskontrolle und Modellverfolgungstools bewältigen die Komplexität verschiedener RAG-Modellversionen und deren Leistung im Laufe der Zeit. Die Einrichtung von Feedback-Schleifen für kontinuierliche Verbesserung, strenge Test- und Validierungsprotokolle sowie das Benchmarking gegenüber Industriestandards gewährleisten hohe Leistung und Zuverlässigkeit. Ethische Richtlinien und Governance werden in MLOps integriert, um Compliance sicherzustellen und Bias zu adressieren. Schulungen für Stakeholder und Change-Management erleichtern die effektive Einführung und Integration von RAG in bestehende Arbeitsabläufe. Die Auswahl der richtigen MLOps-Tools und deren Integration in bestehende Systeme sind entscheidend für die Verbesserung des RAG-Implementierungsprozesses.

Diese Prozesse stellen gemeinsam das effiziente Management, die Bereitstellung und den ethischen Einsatz von KI-Systemen sicher und bieten ein robustes Rahmenwerk, um fortschrittliche KI-Technologien wie RAG effektiv zu nutzen.



Zukünftige Entwicklungen in der Retrieval-Augmented Generation-Technologie

Während Unternehmen weiterhin eine von Daten geprägte Landschaft durchqueren, steht die Weiterentwicklung der RAG-Technologie als Zeugnis der transformativen Kraft der KI. Ein Blick auf die zukünftigen Entwicklungen von RAG zeigt, wie sie die Schnittstelle zwischen Mensch und Information neu gestalten werden.

Verbesserung von Datenabruf und -integration

Die Zukunft von RAG liegt in der weiteren Verbesserung der Datenabrufprozesse. Es werden Fortschritte im Bereich des Natural Language Understanding erwartet, die es RAG-Systemen ermöglichen werden, komplexe und nuancierte Anfragen mit noch größerer Genauigkeit zu interpretieren. Dies könnte zu intuitiveren Interaktionen führen, da RAG-Systeme beginnen, Kontext und Subtext ebenso gut zu verstehen wie ein Mensch.

Fortschritte in der Natural Language Generation

Während generative KI-Modelle weiterhin Fortschritte machen, werden auch die Fähigkeiten von RAG-Systemen zur Erzeugung von reichhaltigem, nuanciertem und vielfältigem Text zunehmen. Die Entwicklung fortschrittlicherer NLG-Komponenten wird es RAG-Systemen ermöglichen, Inhalte zu erstellen, die zunehmend ununterscheidbar von menschlich geschriebenen Texten sind — und das in einem Bruchteil der Zeit.

Integration multimodaler Daten

Die Integration multimodaler Daten stellt eine bedeutende Grenze für die RAG-Technologie dar. Durch die Einbeziehung visueller, auditiver und anderer Datenformen werden RAG-Systeme in der Lage sein, umfassendere Antworten zu liefern, die über die Beschränkungen von Text hinausgehen. Dies verbessert die Benutzererfahrung und eröffnet neue Anwendungsbereiche.

Quantencomputing: Ein neuer Horizont

Quantencomputing bietet eine verlockende Perspektive für RAG-Systeme, mit dem Potenzial, die Geschwindigkeit und Kapazität der Datenverarbeitung exponentiell zu erhöhen. Dies könnte die Effizienz von RAG-Systemen revolutionieren und die Echtzeit-Datenabfrage und -generierung für selbst die komplexesten und umfangreichsten Datensätze ermöglichen.

Unsupervised learning Algorithmen

Neue Unsupervised learning Algorithmen werden voraussichtlich eine bedeutende Rolle in der Zukunft von RAG spielen. Diese Algorithmen ermöglichen es RAG-Systemen, aus Daten zu lernen und sich anzupassen, ohne explizite Programmierung, was zu einem organischeren und weniger ressourcenintensiven Trainingsprozess führt.

Ethische KI: Ein fortlaufendes Engagement

Mit der zunehmenden Verbreitung der RAG-Technologie wird das Engagement für ethische KI immer wichtiger. Zukünftige Entwicklungen in RAG müssen Fairness, Datenschutz und Verantwortlichkeit in den Vordergrund stellen, um sicherzustellen, dass die Vorteile dieser Systeme allen Teilen der Gesellschaft zugutekommen.

Kollaborative KI: Menschen und Maschinen als Partner

Mit Blick auf die Zukunft entsteht ein kollaboratives KI-Ökosystem, in dem RAG-Systeme und menschliche Expertise im Einklang arbeiten. Diese Partnerschaften nutzen die Stärken beider Seiten, wobei Menschen strategische Aufsicht bieten und RAG-Systeme unvergleichliche analytische und generative Fähigkeiten bereitstellen.

Die weltweiten Auswirkungen der RAG-Innovation

Die Auswirkungen von RAG werden global sein, mit potenziellen Anwendungen in vielen Sprachen und Kulturen. Diese weltweite Reichweite erfordert ein tiefes Verständnis der lokalen Kontexte und Gepflogenheiten, um sicherzustellen, dass RAG-Systeme relevante und kulturell sensible Inhalte liefern können.

Abschließende Gedanken: RAG als Katalysator für Wandel

Die zunehmende Verbreitung von Retrieval-Augmented Generation (RAG) stellt einen bedeutenden Fortschritt im Bestreben dar, genauere und Kontext relevantere Antworten zu generieren. Trotz ihres umfangreichen Trainings auf großen Datensätzen haben Large Language Models (LLMs) oft Schwierigkeiten, aktuelle Informationen und proprietäre Daten einzubeziehen. Diese Lücke führt zu den berüchtigten „Halluzinationen", bei denen LLMs mit großer Überzeugung falsche Antworten liefern.

Das Fein-Tuning von LLMs ist eine Strategie, um dieses Problem anzugehen. Laut einer Umfrage von Retool nutzen 29 % der Befragten diesen Ansatz, um die Daten, auf denen LLMs trainiert werden, anzupassen[6]. Allerdings vollzieht sich ein bemerkenswerter Wandel bei größeren Unternehmen. Die gleiche Retool-Umfrage ergab, dass ein Drittel der Unternehmen mit über 5.000 Mitarbeitern nun RAG einsetzt, um auf zeitkritische Daten zuzugreifen, wie beispielsweise Aktienkurse und interne Geschäftsinformationen, wie Kunden- und Transaktionshistorien.

Die Fähigkeit von RAG, Echtzeit-Datenabruf in leistungsstarke generative Modelle zu integrieren, macht es für viele Organisationen zur bevorzugten Methode. Dadurch wird sichergestellt, dass die Antworten nicht nur genau, sondern auch im aktuellsten und relevantesten Kontext verankert sind. Dieser Trend unterstreicht das wachsende Bewusstsein für das Potenzial von RAG, die Lücke zwischen statischem Wissen und dynamischen, realen Daten zu überbrücken, und ebnet den Weg für verlässlichere und effektivere KI-gesteuerte Lösungen.

[6] https://retool.com/reports/state-of-ai-h1-2024

Autoren und Danksagungen

Autoren

Purshottam Purswani
CTO, Atos APAC

Mischa Van Oijen
CTO and Platform Engineering Head

Danksagungen

Die Autoren danken der Atos Research Community (ARC) und insbesondere den folgenden Mitgliedern für ihre wertvollen Kommentare: Gabriel Sala und Erwin Dijkstra.

Über Atos

Atos ist ein weltweit führender Anbieter für die digitale Transformation mit ca. 82.000 Mitarbeitern und einem Jahresumsatz von zirka 10 Milliarden Euro. Als europäischer Marktführer für Cybersecurity sowie Cloud und High Performance Computing bietet die Atos Gruppe maßgeschneiderte, ganzheitliche Lösungen für sämtliche Branchen in 69 Ländern. Als Pionier im Bereich nachhaltiger Dienstleistungen und Produkte arbeitet Atos für seine Kunden an sicheren, dekarbonisierten Digitaltechnologien. Atos ist eine SE (Societas Europaea), die an der Börse Euronext Paris notiert ist.

Das Ziel von Atos ist es, die Zukunft der Informationstechnologie mitzugestalten. Fachwissen und Services von Atos fördern Wissensentwicklung, Bildung sowie Forschung in einer multikulturellen Welt und tragen zu wissenschaftlicher und technologischer Exzellenz bei. Weltweit ermöglicht die Atos Gruppe ihren Kunden und Mitarbeitern sowie der Gesellschaft insgesamt, in einem sicheren Informationsraum nachhaltig zu leben, zu arbeiten und sich zu entwickeln.

Weitere Informationen finden Sie unter www.atos.net

Über Tech Foundations

Tech Foundations umfasst den Managed-Services-Geschäftsbereich der Atos Gruppe mit Fokus auf Hybrid Cloud Infrastructure, Employee Experience und Technology Services. Mit seinen dekarbonisierten, automatisierten und KI-gestützten Lösungen ist Tech Foundations führend in diesem Bereich und treibt mit seinen 41.000 Mitarbeitern Themen voran, die Unternehmen, Institutionen und die Gesellschaft weltweit am dringendsten beschäftigen. Das Unternehmen ist in 69 Ländern vertreten und erzielt einen Jahresumsatz von ca. 5 Milliarden Euro.

Über die Atos Research Community (ARC)

Die ARC ist aus der Atos Scientific Community und der Atos Expert Community hervorgegangen – zwei strategischen Gemeinschaften mit einer reichen Geschichte an Kreativität, Forschung, Vordenkertum und Innovation.

Die ARC ist eine dynamische, vielfältige und inklusive Gemeinschaft von über 1000 Experten, die sich der Weiterentwicklung relevanter Themen für unsere Kunden widmen.

Unser Expertennetzwerk konzentriert sich darauf, neue Technologietrends zu identifizieren und sie auf pragmatische, innovative und agile Weise zur Lösung von Kundenherausforderungen einzusetzen.

Die Ziele der ARC sind ehrgeizig: Grenzen verschieben, wegweisende Lösungen entwickeln und stets einen Schritt voraus sein.

Atos