Inhalt

Einleitung

Entdecken Sie die wahren Kosten eines RAG-Chatbots für Ihr Unternehmen und wie er Ihre digitale Transformation vorantreiben kann! In unserem umfassenden Whitepaper beleuchten wir die finanziellen Aspekte und Vorteile des Einsatzes von RAG-Chatbots. Erfahren Sie, welche Investitionen notwendig sind, welche Einsparungen möglich sind und wie ein RAG-Chatbot den Kundenservice und die Effizienz Ihrer Geschäftsprozesse revolutionieren kann. Lassen Sie sich von konkreten Fallstudien inspirieren und erhalten Sie wertvolle Tipps für eine erfolgreiche Implementierung. Verpassen Sie nicht die Chance, Ihre digitale Transformation auf die nächste Stufe zu heben!

Was ist ein RAG-Chatbot?

Ein RAG-Chatbot (Retrieval-Augmented Generation) ist ein LLM (Large Language Model) basierter Chatbot, der eine Kombination aus Information Retrieval und Natural Language Generation verwendet, um präzise und kontextbezogene Antworten auf Benutzerfragen zu geben.

Funktionsweise

Wenn ein Benutzer eine Frage stellt, durchsucht der RAG-Chatbot zunächst seine Wissensdatenbank (z.B. Dokumente, Webseiten, FAQs), um relevante Informationen abzurufen (Retrieval).

Basierend auf den abgerufenen Informationen generiert ein vortrainiertes LLM dann eine Antwort in natürlicher Sprache, die auf die spezifische Frage des Benutzers zugeschnitten ist (Augmented Generation).

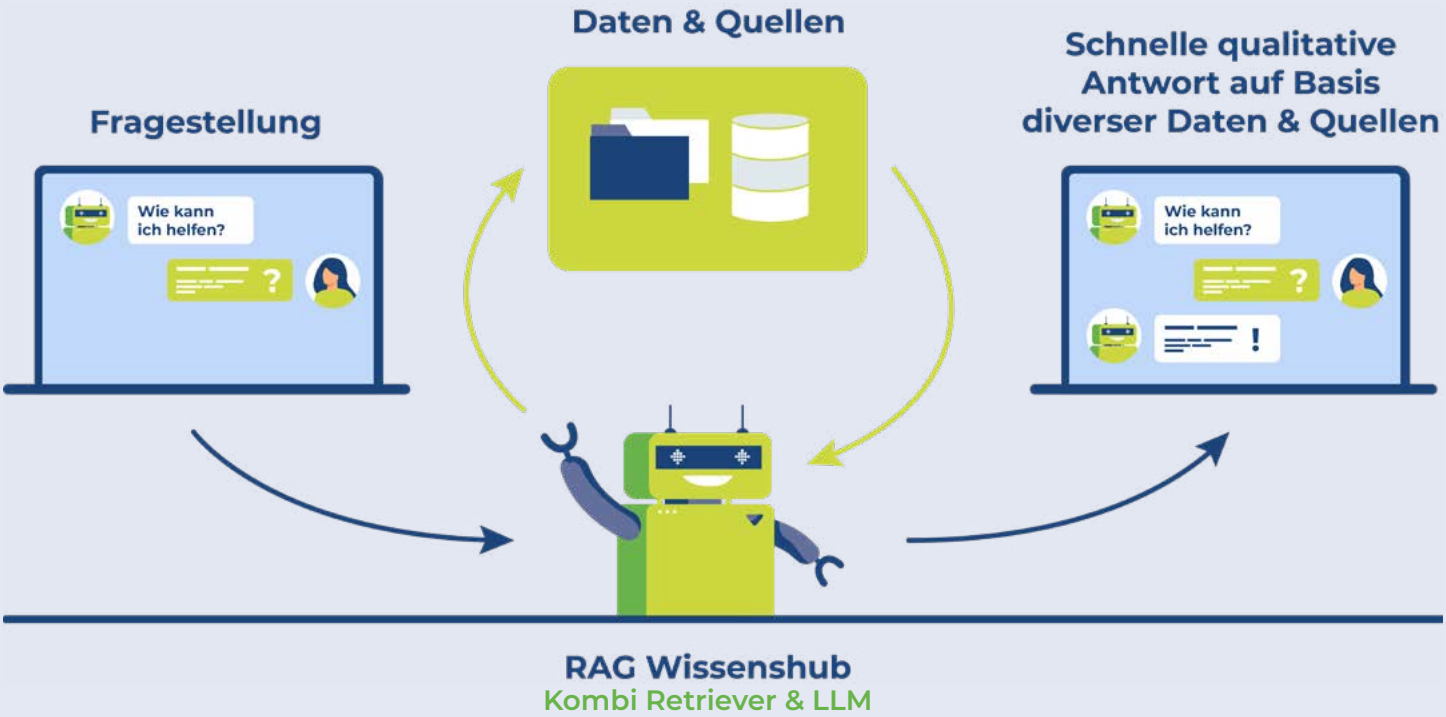
Einsatzbereiche in Unternehmen

Kundensupport: RAG-Chatbots sind in der Lage, rund um die Uhr Kundenanfragen beantworten, Probleme lösen und Informationen bereitstellen, was zu einer höheren Kundenzufriedenheit und Entlastung des Support-Teams führt.

Interne Wissensvermittlung: Mitarbeiter können RAG-Chatbots nutzen, um schnell Antworten auf Fragen zu Unternehmenspolitik, Prozessen oder technischen Themen zu erhalten, ohne zeitaufwendig in Dokumenten suchen zu müssen.

Vertrieb und Marketing: RAG-Chatbots machen es möglich, potenzielle Kunden durch den Verkaufsprozess führen, personalisierte Produktempfehlungen geben und bei der Lead-Generierung unterstützen.

HR und Onboarding: Neue Mitarbeiter können RAG-Chatbots nutzen, um Fragen zu Unternehmenskultur, Benefits oder Einarbeitungsprozessen zu stellen, was den Onboarding-Prozess erleichtert.



Funktionsweise des RAG-Ansatzes

Indem der RAG-Ansatz die externe Wissensbasis einbezieht, ist er in der Lage, präzisere und evidenzbasierte Antworten zu liefern als rein generative Modelle, die externe Wissensquellen nicht nutzen.

Inhaltsvektorisierung: Die Inhalte der Wissensbasis (Dokumente, Webseiten, Datenbanken) werden in numerische Vektoren umgewandelt und in einer Vektordatenbank gespeichert. Dieser Prozess kodiert die semantischen Beziehungen zwischen den Texten.

Anfrage-Vektorisierung: Die Frage oder Anfrage des Benutzers wird ebenfalls in einen Vektor umgewandelt, der die semantische Bedeutung repräsentiert.

Vektorsuche (Retrieval): Der Anfragevektor wird mit den Vektoren in der Vektordatenbank verglichen, um die am besten passenden Kontexte (z.B. Textabschnitte) zu finden.

Kontexterweiterung des Prompts: Die extrahierten Kontexte werden mit dem ursprünglichen Prompt zu einem neuen, erweiterten Prompt für das generative Sprachmodell kombiniert.

Antwortgenerierung: Das generative Sprachmodell (z.B. GPT-3) verarbeitet den kontexterweiterten Prompt und generiert darauf basierend eine natürlichsprachliche Antwort unter Berücksichtigung der relevanten Informationen aus den Kontexten.