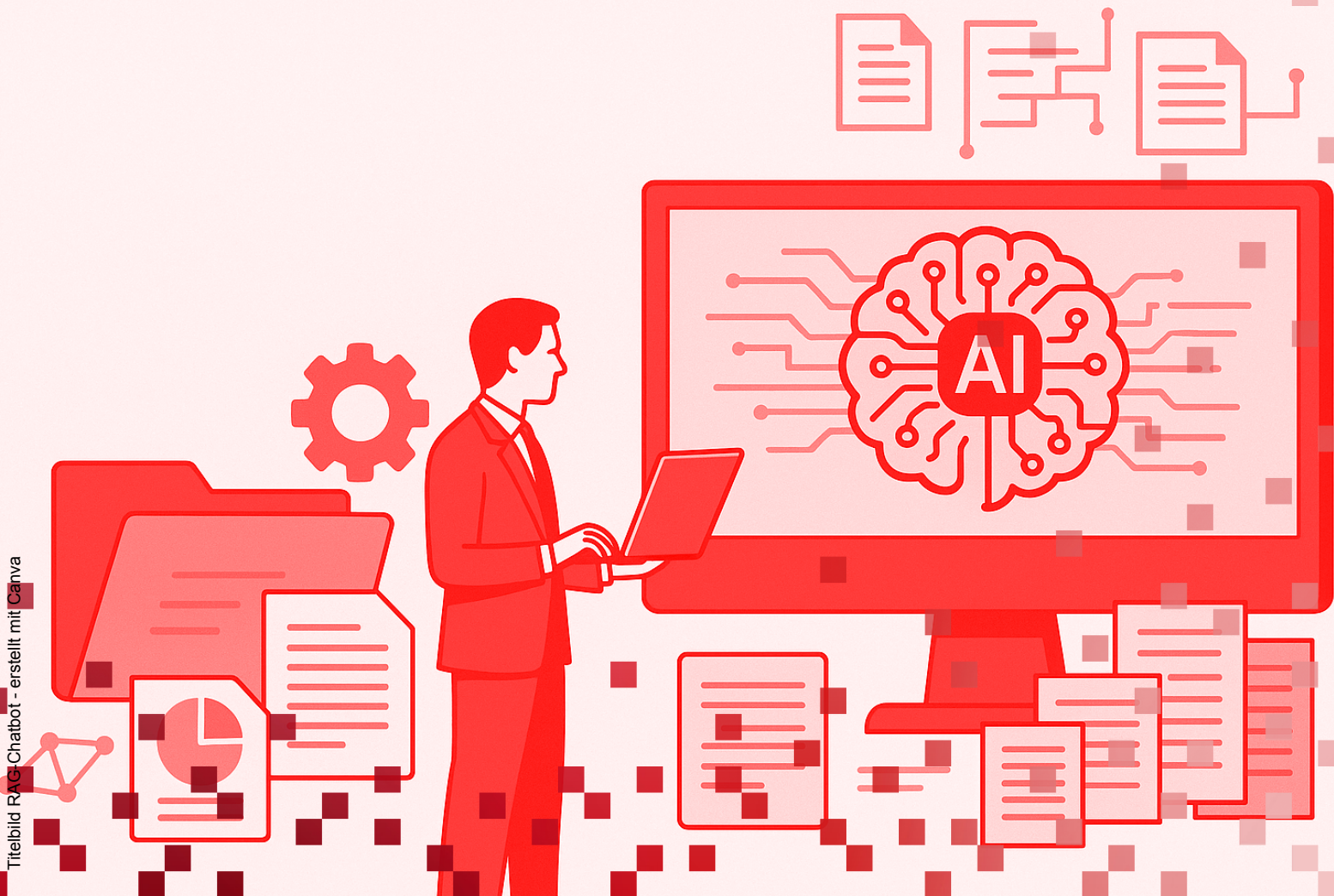




Mittelstand-Digital
Zentrum
WertNetzWerke



Titelbild RAG-Chatbot - erstellt mit Canva

Faktenblatt

Wissenszugang neu gedacht: Chatbots mit Retrieval-Augmented Generation für den Mittelstand

Gefördert durch:



Bundesministerium
für Wirtschaft
und Energie

Mittelstand-
Digital

aufgrund eines Beschlusses
des Deutschen Bundestages

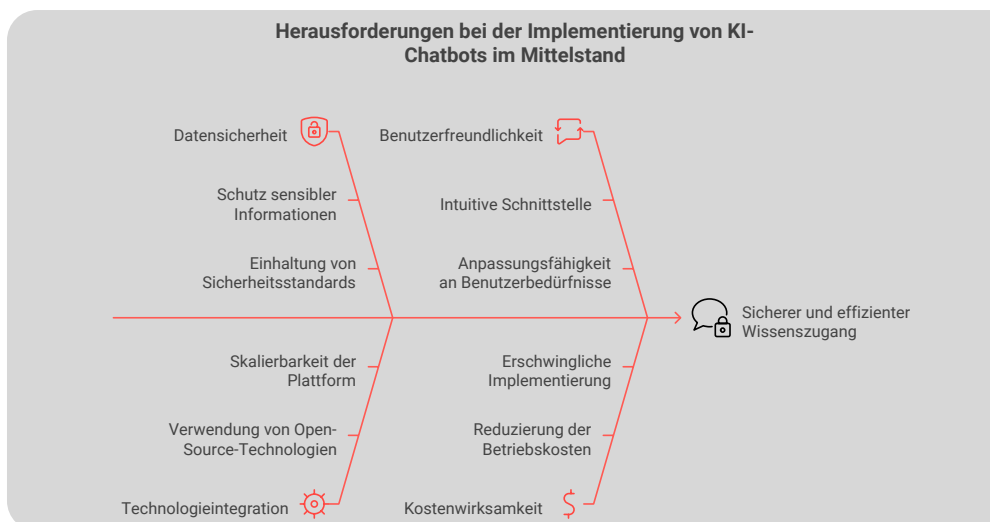
Die Herausforderung, internes Wissen sicher und effizient verfügbar zu machen, kennt jedes mittelständische Unternehmen.

Besonders Unternehmen aus dem Sondermaschinenbau und der Kleinserienfertigung sind auf dieses Wissen angewiesen, um ihre Wettbewerbsfähigkeit zu sichern.

Ein innovativer und produktiver Ansatz besteht darin, generative Künstliche Intelligenz, konkret in einem Chatbot-Format, zu implementieren, um den schnellen Zugriff auf Wissen zu gewährleisten. Dabei ist es entscheidend, dass interne Daten und Dokumentationen nicht für das Training von Sprachmodellen – sogenannten Large-Language-Models (LLMs) – verwendet werden. Der Schutz sensibler Informationen hat dabei höchste Priorität und muss jederzeit sichergestellt sein.

Eine Lösung für dieses Problem bietet die Retrieval-Augmented Generation (RAG). Diese Technologie ermöglicht es, ein vortrainiertes Sprachmodell so einzusetzen, dass bei jeder Anfrage auf interne, firmeneigene Daten zugegriffen wird, ohne diese für das Training des Modells zu verwenden. Damit bleibt das Wissen innerhalb des Unternehmens geschützt, während der Chatbot dennoch in der Lage ist, präzise Antworten zu liefern. Die Kombination aus Sicherheit und Anpassungsfähigkeit bietet genau die Eigenschaften, die sich im Mittelstand als besonders wirkungsvoll und praxisnah erwiesen haben.

Erste Anwendungen auf vorhandenen Datenbeständen zeigen das Potenzial der Lösung: Eine sichere und skalierbare Plattform ermöglicht es, Daten effizient und unter Einhaltung höchster Sicherheitsstandards zu nutzen, ohne diese nach außen preiszugeben. Basierend auf diesen Prinzipien lässt sich zum Beispiel auch eine umfassende, modulare Dienstleistung entwickeln, die auf verschiedenen bewährten Open-Source-Technologien aufbaut:



1. Plattformen, die auf Open-Source-Lösungen setzen:

- **Hugging Face Transformers:** Eine Open-Source-Bibliothek, die eine Vielzahl vortrainierter Modelle bietet und die es ermöglicht, diese individuell an die Anforderungen des Unternehmens anzupassen.
- **GPT-J und GPT-NeoX von EleutherAI:** Diese Open-Source-Sprachmodelle sind eine leistungsstarke Alternative zu proprietären Lösungen und können flexibel sowohl lokal als auch in der Cloud betrieben werden.
- **BLOOM:** Ein Open-Source-Sprachmodell, das von der BigScience-Initiative entwickelt wurde und für viele verschiedene Sprachverarbeitungsaufgaben geeignet ist.

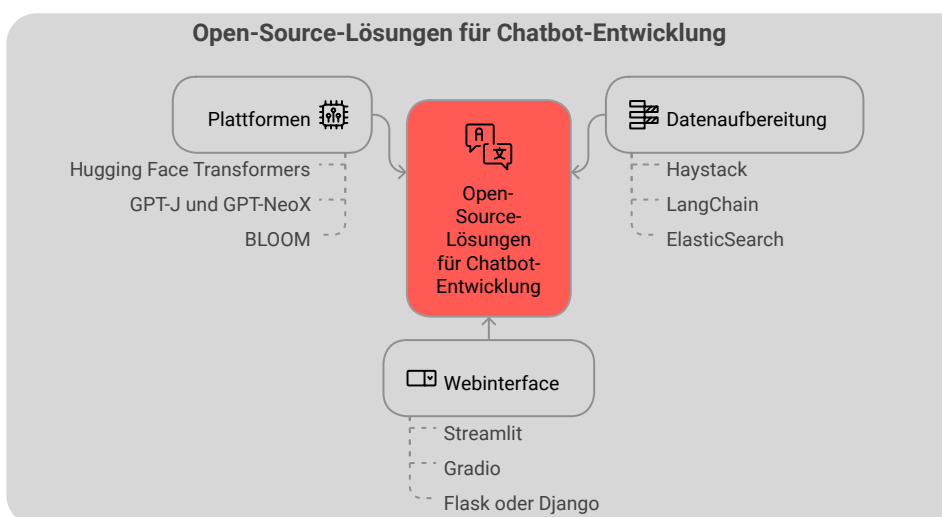
2. Datenaufbereitung und -integration:

- **Haystack** von deepset.ai: Ein Open-Source-Framework, das speziell für die Umsetzung von Retrieval-Augmented Generation entwickelt wurde und die Entwicklung von leistungsfähigen Such- und Antwortsystemen unterstützt.
- **LangChain:** Eine Open-Source-Bibliothek, die Large Language Models mit externen Datenquellen verbindet und Werkzeuge für die maßgeschneiderte Datenaufbereitung bereitstellt.
- **ElasticSearch:** Eine Open-Source-Such- und Analyse-Engine, die es ermöglicht, große Datenmengen schnell und effizient zu durchsuchen und zu analysieren.

3. Ein benutzerfreundliches Webinterface:

Für die Interaktion mit dem Chatbot wird ein Webinterface auf Basis von Open-Source-Tools entwickelt:

- **Streamlit:** Ein Open-Source-Framework, das die schnelle und einfache Erstellung interaktiver Webanwendungen ermöglicht, die sich ideal für Prototypen und interne Tools eignen.
- **Gradio:** Ein Open-Source-Tool, das eine benutzerfreundliche Erstellung von Webinterfaces für Machine-Learning-Modelle mit minimalem Codeaufwand ermöglicht.
- Alternativ können auch **Flask** oder **Django** verwendet werden, beide etablierte Open-Source-Webframeworks.



In mittelständischen Unternehmen ist die effiziente Nutzung von internem Wissen oft entscheidend, insbesondere wenn Informationen über verschiedene Dokumentationssysteme verteilt sind. Ein Chatbot, der auf RAG basiert, kann hier einen strukturierten Zugang zu internem Wissen bieten, ohne die Vertraulichkeit der Daten zu gefährden.

Durch die Kombination dieser Technologien eröffnet die RAG-basierte Lösung neue Möglichkeiten nicht nur für die internen Prozesse, sondern auch für die Vermarktung komplexer Produkte. Potenzielle Kund:innen profitieren von der Möglichkeit, über einen Chatbot schnell und unkompliziert Informationen zu komplexen Produkten abzurufen, was zu einer deutlichen Beschleunigung des Verkaufsprozesses führen kann. Gleichzeitig eröffnet der Einsatz solcher Systeme Mitarbeitenden die Möglichkeit, gezielte Anfragen zu stellen und unmittelbar relevante Informationen aus firmeninternen Datenquellen wie Datenbanken, Handbüchern oder technischen Dokumentationen zu erhalten. Insbesondere im Außendienst kann dies die Lösung von Problemen beschleunigen, indem der Chatbot schnell Zugang zu detaillierten Anweisungen und Problemlösungen bietet. Der Einsatzzweck hängt einzig nur von der Art der Daten ab, die zum Training verwendet wurden.

Mit der Einführung von Retrieval-Augmented Generation gelingt es, internes Wissen sicher und effizient nutzbar zu machen. Diese Lösung basiert auf erprobten Open-Source-Technologien und bietet dem Mittelstand die Möglichkeit, sowohl interne Prozesse zu optimieren als auch neue Marketingchancen zu nutzen. So schafft die innovative Technologie einen echten Mehrwert für Unternehmen, die auf der Suche nach modernen, zukunftsfähigen Lösungen sind.

Impressum

Herausgeber:
Mittelstand-Digital Zentrum
WertNetzWerke
wisnet innovation research
institute e. V.
c/o HAGEN.WIRTSCHAFTS-
ENTWICKLUNG GmbH
Rathausstr. 2, 58095 Hagen

hagen@wertnetzwerke.de
www.wertnetzwerke.de

Autor:
Ralph Jan Wörheide

Redaktion:
Nico Piepenstock

Stand:
September 2025