RPA-Blog

16.05.2024

Mit Retrieval-Augmented Generation zum effizienten Workflow

Chatbots wie ChatGPT sind seit geraumer Zeit in aller Munde. Basierend auf Künstlicher Intelligenz (KI) erstellen sie Texte, Bilder oder Videos in Sekundenschnelle und in oftmals beeindruckender Qualität. Insbesondere mit ihrer Fähigkeit, individuelle Fragen zu interpretieren und zu beantworten, erwecken sie den Eindruck, als säße ein Mensch am anderen Ende der Leitung.

Natürlich stellen sich da viele Unternehmen die Frage, ob und wie sie solche KI-Technologien einsetzen können, um ihre Workflows zu straffen. Von der Kundenbetreuung über die Fehlerprüfung bis zur Datenbankabfrage bietet die scheinbare Intelligenz schließlich unendliche Möglichkeiten zur Optimierung.

Der Teufel steckt im Detail

Um zu verstehen, warum das nicht ganz so einfach ist, braucht es einen Blick hinter die Kulissen: Chatbots wie ChatGPT funktionieren auf Grundlage sogenannter generativer KI-Modelle. Diese werden zunächst mit gewaltigen Datenmengen trainiert. Alles, was der Chatbot ausgibt, basiert auf diesen Informationen. Zwar können die Bots extrapolieren, also sozusagen eigene Schlüsse aus dem bestehenden Datenmaterial ziehen. Ihre Antworten gehen aber niemals über das hinaus, womit sie ursprünglich „gefüttert“ worden sind.

Wenn die zugrundliegenden Daten veraltet oder gar fehlerhaft sind, kann der Bot keine oder schlimmstenfalls eine falsche Antwort liefern. Bei ChatGPT wurde beispielsweise gezeigt, dass die Software fehlende Informationen in vielen Fällen einfach erfindet – Fachleute sprechen dann von „Halluzinationen“. Hinzu kommt, dass der Bot häufig keine Quellen für seine Ergebnisse liefern kann, womit die Antworten kaum überprüft werden können.

Was bei einer vom Chatbot erstellten Hausaufgabe noch für ausgelassene Heiterkeit im Klassenzimmer sorgen kann, stellt Unternehmen vor gewaltige Probleme. Denn sie müssen sich bei Kundenanfragen oder internen Datenbanksuchen unbedingt auf die Richtigkeit der Angaben verlassen können.

Generative KI-Modelle können das alleine nicht leisten, da sie nicht mit den spezifischen Unternehmensdaten trainiert worden sind. Und selbst wenn sie es wären, könnte die KI Veränderungen der Daten nur mit aufwendigen und teuren Trainings miteinbeziehen.

Retrieval-Augmented Generation: Das Beste aus zwei Welten

Mit Retrieval-Augmented Generation (RAG) lässt sich diese Schwachstelle beheben. Die Technologie ermöglicht es der generativen KI, ihre statischen Datengrundlagen mit aktuellen, dynamischen

Informationen anzureichern. Sie kann dann verschiedenste Quellen wie beispielsweise Dokumente, Datenbanken oder Webseiten für ihre Antworten berücksichtigen, ohne dass ein erneutes Training nötig wird.

Gleichzeitig behält der Chatbot die Fähigkeit zur semantischen Suche – er kann also auf komplexe Fragestellungen kontextbezogen und verständlich antworten. Damit kommen Mitarbeiter und Führungskräfte in Sekundenschnelle an die benötigten Informationen, ohne sich über Stichworte oder Datenbankstrukturen Gedanken machen zu müssen. RAG bezieht dabei sogar frühere Anfragen mit ein, die KI lernt also stetig dazu.

Auch bei der Auswertung des Datenmaterials kann RAG glänzen, wie Forscher von Facebook AI Research, des University College London und der New York University herausgefunden haben. „Dokumente, die Hinweise auf die Antwort geben, aber die Antwort nicht wortwörtlich enthalten, können dennoch zur Generierung einer korrekten Antwort beitragen, was mit herkömmlichen […] Ansätzen nicht möglich ist“, so ihr Fazit. „RAG kann richtige Antworten generieren, auch wenn die richtige Antwort in keinem der abgerufenen Dokumente enthalten ist.“

Hinzu kommt, dass jede Antwort der KI mit einer Quellenangabe versehen werden kann. Damit ist sichergestellt, dass der Bot nicht unbemerkt „halluziniert“. Zudem können Fehler in Dokumenten oder Ähnlichem auf diesem Weg direkt identifiziert und verbessert werden.

Der digitale Assistent „Fred“

„Fred“ ist ein von Brain4Data entwickelter digitaler Assistent, der Fähigkeiten aus den Bereichen Robotic Process Automation (RPA) und Augmented Intelligence (AI) vereint. Er optimiert Arbeitsabläufe, indem er Informationen bündelt, aufbereitet und zum richtigen Zeitpunkt zur Verfügung stellt. Damit entlastet „Fred“ seine menschlichen Kollegen und reduziert unnötigen Kommunikationsaufwand im Unternehmen.

Der digitale Assistent nutzt Retrieval-Augmented Generation (RAG) und generative KI-Modelle der Oracle Cloud Infrastructure. Er arbeitet abteilungsübergreifend und verknüpft so alle relevanten Informationen – beispielsweise aus Vertrieb, Produktion, Logistik und Buchhaltung.

Der digitale Assistent warnt selbstständig bei wichtigen Geschäftsvorfällen und liefert maßgeschneiderte Handlungsempfehlungen. Daneben besitzt der „Fred“ Module für die Vertriebsoptimierung, Projekt- und Personaleinsatzplanung, Bauobjektqualifizierung und einiges anderes mehr. Der digitale Assistent kann Daten aus allen gängigen Programmen wie etwa Excel oder SAP-Anwendungen verarbeiten. Das bedeutet, dass Unternehmen ihre bestehende IT-Infrastruktur nicht verändern müssen und Mitarbeiter mit der gewohnten Software weiterarbeiten können.

Über Brain4Data

Die Brain4Data GmbH und Co. KG entwickelt in Saarwellingen Lösungen in den Bereichen Robotic Process Automation (RPA) und Augmented Intelligence (AI). Wir analysieren Kommunikations- und Arbeitsprozesse in Unternehmen. Auf dieser Basis identifizieren wir Automatisierungspotenziale, die einen reibungslosen Arbeits- und Kommunikationsfluss unterstützen. Darüber hinaus bieten wir den

digitalen Assistenten „Fred“ an, mit dessen Hilfe Prozesse automatisiert und Arbeitsabläufe vereinfacht werden können.

Weitere Informationen zur Brain4Data und dem digitalen Assistenten „Fred“ unter:

https://brain4data.de/

Kontakt

David Woirgardt-Seel

Chief Knowledge Officer

Telefon: +49 (0) 6838 502 09 63

E-Mail: david.seel@brain4data.de