Dem ChatBot das Halluzinieren austreiben

Retrieval Augmented Generation ist die nächste Entwicklungsstufe Stufe künstlicher Intelligenz. Sie kombiniert Textgeneratoren mit Information-Retrieval aus externen Wissensquellen.

VON MATTHIAS BAUER

der künstlichen Intelligenz (KI), das Daten aus externen Wissensquellen abruft, um die Qualität der Ergebnisse einer genAI zu verbessern. Große Sprachmodelle (Large Language Models = LLM) wie ChatGPT oder Bard besitzen kein Wissen im eigentlichen Sinn, sondern können vor allem gut von Input auf Output schließen. Das von einem LLM aus seinen Trainingsdaten gelernte Wissen bezeichnet man als parametrisches Gedächtnis. Es wird in seinen neuronalen Gewichten (einer Art numerischer Repräsentation) gespeichert. LLMs antworten auf Abfragen aus dem parametrischen Speicher. In den meisten Fällen aber ist die Informationsquelle unbekannt und ein LLM kann keine wörtlichen Zitate liefern.

Die Ergebnisse waren mehr als erstaunlich: Antworten lassen sich in Abhängigkeit der jeweils anzusprechenden Zielgruppe unterschiedlich formulieren und dann auch sofort nutzen. Hinzu kommt eine hohe Transparenz durch **Angabe der Quellen**.

GROSSE SPRACHMODELLE KENNEN auf fast alle Fragen eine Antwort. Nicht immer die richtige (Stichwort Bullshit-Generator), nicht immer die Aktuellste (ChatGPT endet gegenwärtig 2021). Beeindruckend ist die neue Technologie allemal und sie gewinnt noch an Faszination, wenn man die Skills des Chatbots mit eigenen Informationen verbindet – mit allem also, was in den unternehmenseigenen Anwendungen an Wissen steckt. Retrieval Augmented Generation (RAG) heißt das Schlagwort: ein Verfahren

Keine Halluzinationen mehr

Die Sprachmodelle sind deshalb gut im Generieren von Inhalten aus einem allgemeinen Korpus ihrer selbst heraus. Unternehmen müssen aber Text aus privaten Wissensdatenbanken erzeugen, das heißt aus ihrem Intranet, Sharepoint- und Confluence-Seiten, PDFs, Office-Dokumenten und so

DER AUTOR
Matthias Bauer ist Teamlead Data Science bei X-Integrate.

weiter. Auf der Basis dieser Wissensdatenbanken ist es dann die Aufgabe von KI-Assistenten, genaue und relevante Antworten zu liefern. Damit umgeht man das derzeit noch größte Manko generativer Chatbots: ihre notorische Neigung zum Halluzinieren. „Prompten" nennt man die Eingabeaufforderung im genAI-Sprech. Der an das LLM gesendete Prompt wird beim beschriebenen Verfahren durch relevante Daten ergänzt, die über einen Information-Retrieval-Mechanismus aus externen Wissensdatenbanken (dem nicht-parametrischen Gedächtnis) abgerufen werden. Diese Daten werden als Kontext zusammen mit der Frage verwendet und der parametrische Speicher damit ausgespart.

Vektor- statt indexbasierte Datenbanken

Der Prompt wird umgewandelt in numerische Werte (wofür man ebenfalls ein Sprachmodell benötigt, ein sogenanntes Embedding Model) und an eine Vektor- oder Graphdatenbank geleitet, in der sich das gesamte Unternehmenswissen befindet. Mittels der numerischen Werte werden dann in der Datenbank die entsprechenden Einträge gefunden, wieder in Text umgewandelt sowie an das LLM gegeben, und zwar mitsamt ihrer Metadaten(z.B. Ablageort und Ersteller eines PDFs). Damit enthält die Antwort zugleich die Quelle. Dies ist gegenwärtig die beste Methode, um Content zu finden; feingetunt oder trainiert werden müssen die LLMs oft nicht – es kann aber in manchen Use Cases nützlich sein. Voraussetzung dafür ist es, den gesamten Content des Unternehmens permanent zu vektorisieren. Dafür müssen bestehende Datensilos im Unternehmen aufgebrochen und zusammengebracht werden. Klassische Unternehmenssuche arbeitet bislang vor allem mit Indexdatenbanken, die mitunter nicht die gewünschte Antwortqualität liefern.

Generative KI-Systeme in die Gegenwart holen

Man spricht bei diesem Verfahren von Retrieval Augmented Generation. RAG braucht man in allen Bereichen, in denen es auf viele aktuelle und genaue Informationen ankommt. Also im Prinzip überall. Durch die Verwendung aktueller und kontextspezifischer Daten holt RAG generative KI-Systeme in die Gegenwart. Die Technik liefert deutlich genauere Ergebnisse auf Anfragen als ein generatives großes Sprachmodell allein.

In RAG liegt daher die Zukunft der Unternehmenssuche. Das Ganze ist keine Utopie, sondern funktioniert schon heute, wie X-Integrate in einer Reihe von Projekten im Finanzdienstleistungssektor erprobt hat. Eine Versicherung zum Beispiel hat mittels RAG einen Bot programmiert, der auf Sachstandsanfragen aus E-Mails oder Call-Center-Ereignissen in angepasstem Kommunikationsstil der Versicherung automatisiert und individuell antwortet. Oder die Bank, die Finanzprodukte auf ihr Risiko hin bewerten muss. Der Risiko-Score wird gebildet auf der Basis regulatorischer Anforderungen, interner Verfahrensanweisungen und Compliance-Richtlinien, die in Tausenden von Dokumenten stecken. Kein Controller kann all diese Informationen überblicken.

Die Bank hat sie vollständig vektorisiert und lässt die Prompts mit Fragen zur Risikoeinschätzung nun auf diese Vektordatenbank los. Die Ergebnisse waren mehr als erstaunlich, einerseits hinsichtlich der Exaktheit bzw. Akkuratesse der Antworten, andererseits hinsichtlich der Verständlichkeit der Sprache: Antworten lassen sich in Abhängigkeit der jeweils anzusprechenden Zielgruppe unterschiedlich formulieren und dann auch sofort nutzen. Hinzu kommt eine hohe Transparenz durch Angabe der Quellen. Denn wo der Ursprungstext mit der jeweiligen Antwort eigentlich herkommt, darüber schweigt sich ChatGPT zum gegenwärtigen Zeitpunkt bekanntlich noch aus. Insgesamt also Resultate, die erahnen lassen, wie groß das Anwendungsfeld für RAG in der näheren Zukunft noch ist. •