PRESSEINFORMATION

Erster Anwendungsfall: EU-Maschinenverordnung

Retrieval Augmented Generation macht das Lesen dicker Wälzer überflüssig

Wer kennt das nicht? Ob mehrere hundert Seiten starke Bedienungsanleitungen für Autos oder umfangreiche Rechtstexte – die entscheidende(n) Stelle(n) zu finden, ist ganz schön zeitaufwendig. Und hat man auch wirklich alle relevanten Einträge gefunden? Ein Register, das den gesuchten Begriff nicht kennt, und unzählige Querverweise machen die Recherche nicht leichter. Das könnte bald der Vergangenheit angehören: Ein Team am Fraunhofer IWU hilft nun einer KI-Lösung auf die Sprünge, »füttert« sie mit umfangreichen Fachtexten und bereitet den externen Input so auf, dass Suchanfragen (Prompts) zu präzisen und erschöpfenden Informationen führen. Retrieval Augmented Generation (RAG) macht's möglich.

Large Language Models (LLMs, große Sprachmodelle) liefern in Chatbots auf alltägliche Anfragen häufig gute bis sehr gute Ergebnisse. Für eine erste Orientierung sind die Antworten meist mehr als ausreichend. Man sollte jedoch auch ihre Grenzen kennen: Trainingsdatensätze können unvollständig oder veraltet sein, einige Informationen unscharf oder gar falsch. Eine Überprüfung der erhaltenen Auskünfte ist daher ratsam. Geht es um Rechtsfragen, sollte man sich nicht ›blind‹ auf einen Chatbot verlassen. Was also tun, wenn von den Angaben die Sicherheit von Menschen abhängen kann? Doch wieder selbst eingehend relevante Dokumente studieren?

RAG sorgt für verlässliche Aussagen zur EU-Maschinenverordnung

Das muss nicht sein – wenn dem Sprachmodell mittels RAG zusätzliche Leitplanken eingezogen werden. Es durchleuchtet dann in erster Linie maßgebliche Texte bzw. Textstellen. Das Sprachmodell wird dabei nicht neu trainiert, sondern selektiv erweitert. Beispielhaft baut das Fraunhofer IWU nun die Maschinenverordnung der EU (2023/1230) in ein LLM ein.

Betrieb soll auch auf Standard-PCs möglich sein

Die Wahl für ein geeignetes Modell fiel dabei auf LLaMA (Large Language Model Meta AI), das groß bzw. leistungsfähig genug ist und dennoch Rechenleistung und Grafikkarte eines hochwertigen Standard-PCs nicht überfordert. Die Anwendung kann somit auf einem lokalen Rechner betrieben werden. In diesem Fall behalten die

Kontakt Pressestelle
Andreas Hemmerle | Fraunhofer-IWU | Telefon +49 371 5397-1372 |
Reichenhainer Straße 88 | 09126 Chemnitz | www.iwu.fraunhofer.de | presse@iwu.fraunhofer.de |

Unternehmen die vollständige Hoheit über ihre Daten. Bei weniger sensiblen Daten ist auch der Betrieb in der Cloud möglich.

So funktioniert RAG: in mehreren Schritten zu faktenbasierten Antworten

Zunächst müssen die ins LLM zu importierenden Daten auf den reinen Text reduziert werden (Cleaning). Sobald dieser in kleinere Abschnitte (Chunks; auffindbare Bausteine) segmentiert ist, geht es mit dem Aufbau eines Suchsystems (Retrieval System) weiter, das die Chunks effizient durchsuchen kann. Die Chunks werden nach relevanten Passagen gegliedert und in einer Vektordatenbank abgelegt, also in mathematische Vektoren umgewandelt, die ihre Bedeutung repräsentieren. Auch die Prompts werden in Vektoren umgewandelt. So wird das Modell in die Lage versetzt, nach den in der Anfrage enthaltenen Wörtern zu suchen und gleichzeitig den Prompt tatsächlich zu ›verstehen‹ (semantische Suche). Das Modell kann jetzt die zu einer Nutzeranfrage passenden Chunks kürzen, neu strukturieren, die wichtigsten Informationen herausfiltern und zu einem verständlichen Zusammenhang kombinieren. Liegt eine konkrete Suchanfrage vor, stehen ausgewählte Chunks zur Verfügung, auf deren Grundlage das Modell faktenbasierte Antworten geben kann. Das Modell nutzt den zusätzlichen Kontext der Chunks und muss nicht neu trainiert werden.

Die richtige Strukturierung macht den Unterschied

Werkzeugmaschinen zählen zu den Kernkompetenzen des IWU. Das Team ›Maschinelles Lernen in der Produktion‹ weiß, auf welche Filter und Vorstrukturierungen es ankommt, damit die jeweils entscheidenden Passagen der Maschinenverordnung erreicht werden. Ein wichtiges Augenmerk liegt künftig außerdem auf weiteren Datenquellen wie Tabellen oder Bildern, die so integriert werden müssen, dass sie gut auffindbar sind.

Smartes KI-Tool für Rechtstexte, Bedienungsanleitungen, Verfahrensanweisungen…

Die Maschinenverordnung soll für einheitliche Standards innerhalb der EU sorgen. Sie definiert beispielsweise, wann bei Änderungen an Maschinen und Anlagen eine neue Konformitätsbewertung erforderlich ist. Als Beispiel für einen anspruchsvollen Rechtstext macht sie den Anfang bei den Demonstrationsanwendungen am IWU. Solche Texte müssen kleine oder mittlere Unternehmen, die über keine größeren Expertenstäbe verfügen können, künftig nicht mehr schrecken: Das Chemnitzer Expertenteam bietet seine Methodenkompetenz an, um mit interessierten Firmen maßgeschneiderte Anwendungen aufzubauen. Infrage kommen dafür Bedienungsanleitungen, die Angebotserstellung oder die Automatisierung von Programmierungsarbeiten im Produktionskontext ebenso wie ›lästige‹ Berichtspflichten.

Das Durchsuchen betriebsinterner Dokumente, Regelungen, Betriebsvereinbarungen oder operativer Daten kann künftig ebenfalls KI-gestützt erfolgen.

Abb. 1 Die Qualität der Antworten von Chatbots, die auf großen Sprachmodellen basieren, hängt von den Trainingsdaten ab: Grundlage sind meist eine Fülle öffentlich zugänglicher Dokumente. Die Verlässlichkeit der Antworten sollten Nutzerinnen und Nutzer kritisch prüfen. Symbolbild: iStock/Galeanu Mihai

Abb. 2 Retrieval Augmented Generation (RAG) nimmt große Sprachmodelle – bildlich gesprochen – an die Hand und sorgt für faktenbasierte Antworten. Bild generiert mit KI (Adobe Firefly)