Fraunhofer IWU

FRAUNHOFER IWU

PRESS RELEASE

It starts with the EU Machinery Regulation

Retrieval Augmented Generation Makes Reading Thick Volumes Obsolete

Who does not know the struggle? Whether several hundred pages of car manuals or extensive legal texts, picking out crucial information is often time-consuming. And will you come across all the relevant entries? A glossary that does not know the search term and countless cross-references make things even more difficult. Such trouble could soon be a thing of the past: A team at Fraunhofer IWU is now helping an AI solution by "feeding" it with extensive technical and legal texts. The team prepares the external input so that search queries (prompts) lead to precise and exhaustive information. Retrieval Augmented Generation (RAG) makes this possible.

Large Language Models (LLMs) often deliver decent results for everyday queries in chatbots. For initial orientation, the answers are usually more than sufficient. However, one should also be aware of their limitations: Training datasets may be incomplete or outdated, and some information may be vague or incorrect. Therefore, verifying the received information is advisable. Regarding legal matters, one should not rely unreservedly on a chatbot. So, what should you do when safety might depend on the information provided? Should you go back to studying relevant documents in detail?

RAG Ensures Reliable Statements on EU Machinery Regulation

That need not be the case when RAG provides additional guidelines to the language model. The model then primarily scans essential texts or text sections. For this purpose, the language model is not retrained but selectively expanded. Fraunhofer IWU is now integrating the EU Machinery Regulation (2023/1230) into an LLM for demonstration purposes.

Operations Also Possible on Standard PCs

The team opted for LLaMA (Large Language Model Meta AI) as an appropriate model, as it is large and powerful enough yet does not overwhelm the computing power and graphics card of a high-quality standard PC. A local machine means that companies retain control over their data. When companies process data with less or no criticality, cloud operation may also be a good choice.

Contact
Andreas Hemmerle | Fraunhofer IWU | Phone +49 371 5397-1372 |
Reichenhainer Straße 88 | 09126 Chemnitz | Germany | www.iwu.fraunhofer.de | andreas.hemmerle@iwu.fraunhofer.de |

How RAG Works: In Several Steps to Fact-based Answers

First, the data selected for import into the LLM needs to be reduced to plain text (Cleaning). Second, the cleaned text is segmented into smaller sections (chunks; discoverable units). Step 3 is to build a search system (Retrieval System) that can efficiently search these chunks. The chunks are organized by relevant passages and stored in a vector database, i.e., converted into mathematical vectors representing their meaning. Prompts are also converted into vectors. This way, the model can search for the words in the request and simultaneously "understand" the prompt (semantic search). The model can now shorten, restructure, extract the most relevant information, and combine it into a comprehensible context. When users enter a specific search query, selected chunks are available, and the model can provide complete, fact-based answers. The model learns from the contexts of the chunks and needs no retraining.

The Right Structuring Makes the Difference

Machine tools are among the core competencies of IWU. The team from the 'Machine Learning in Production' department knows what filters and pre-structuring are essential to reach the critical passages of the Machinery Regulation. A primary concern for the future is the integration of diverse data sources, including tables and images, to make them easily discoverable.

Innovative AI Tool for Legal Texts, Manuals, Standard Operating Procedures...

The Machinery Regulation aims to ensure uniform standards within the EU. For example, it defines when changes to machines and equipment require a new conformity assessment. As an example of a complex legal text, it was the starting point for demonstration applications at IWU. In the future, small and medium-sized enterprises without large expert teams will no longer need to fear such texts: The Chemnitz-based expert team offers its methodological expertise to help interested companies build customized applications. This offer includes manuals, preparing offers, automating programming tasks in production, or strenuous reporting obligations. Manually searching through internal documents, regulations, company agreements, or operational data should also belong to the past soon.

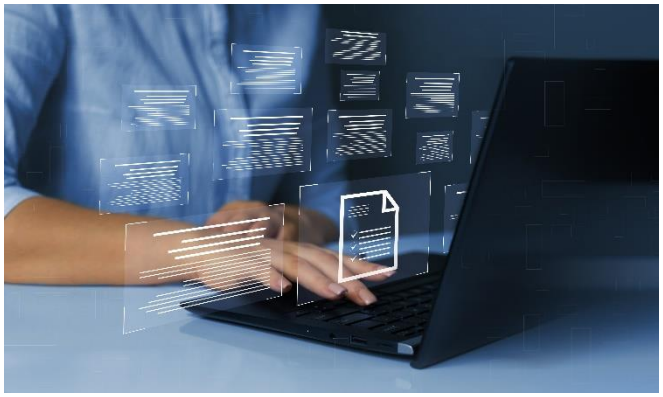

Fig. 1 The quality of answers from chatbots based on large language models depends on the training data, often a plethora of publicly available documents. Users should critically assess the reliability of the answers. Symbol image: iStock/Galeanu Mihai

Fig. 2 Retrieval Augmented Generation (RAG) supports large language models, ensuring fact-based answers. Image generated with AI (Adobe Firefly)