RPA-Blog

16.05.2024

How Retrieval Augmented Generation ensures an efficient workflow

Chatbots such as ChatGPT have been a hot topic for some time now. Based on Artificial Intelligence (AI), they produce texts, images or videos in a matter of seconds and often in impressive quality. Especially with their ability to interpret and answer individual questions, they give the impression that a person is sitting at the other end of the line.

Naturally, many companies are asking themselves whether and how they can use such AI technologies to streamline their workflows. After all, from customer service to troubleshooting and database queries, the supposed intelligence offers endless possibilities for optimization.

The devil's in the details

To understand why this is not quite so simple, we need to take a look behind the scenes: Chatbots such as ChatGPT work on the principle of generative AI models. These are initially trained with huge amounts of data. Everything the chatbot puts out is based on this information. The bots can extrapolate, i.e. draw their own conclusions from the existing data. However, their answers never go beyond what they were originally "fed".

If the underlying data is outdated or even incorrect, the bot can provide no answer or, in the worst case, an incorrect answer. In the case of ChatGPT, for example, it has been shown that in many cases the software simply makes up missing information – experts refer to this as "hallucinations". In addition, the bot is often unable to provide any sources for its results, making it almost impossible to verify the answers.

What can bring laughter to the classroom when a homework assignment is generated by a chatbot poses huge problems for companies. They must be able to rely on the accuracy of the information when dealing with customer inquiries or internal database searches.

Generative AI models cannot achieve this on their own, as they have not been trained with the specific company data. And even if they were, the AI could only incorporate changes to the data with complex and expensive training.

Retrieval-Augmented Generation: The best of both worlds

Retrieval Augmented Generation (RAG) can eliminate this weakness. The technology enables generative AI to enrich its static data bases with up-to-date, dynamic information. It can then incorporate a wide variety of sources such as documents, databases or websites into its answers without the need for retraining.

At the same time, the chatbot retains the ability to perform semantic searches, meaning it can respond to complex questions in a contextual and comprehensible manner. This means that employees and managers can access the information they need in a fraction of a second without having to think about keywords or database structures. RAG even takes previous queries into account, so the AI is constantly learning.

RAG can also excel at evaluating data material, as researchers from Facebook AI Research, University College London and New York University have discovered. "Documents with clues about the answer but do not contain the answer verbatim can still contribute towards a correct answer being generated, which is not possible with standard extractive approaches, they conclude. "Furthermore, RAG can generate correct answers even when the correct answer is not in any retrieved document."

In addition, every answer from the AI can be tagged with a source. This ensures that the bot does not "hallucinate" without being noticed. In that way, errors in documents or the like can also be directly identified and rectified.

"Fred", your digital assistant

"Fred" is a digital assistant developed by Brain4Data who combines skills from the fields of Robotic Process Automation (RPA) and Augmented Intelligence (AI). He streamlines work processes by bundling and processing information and making it available at the appropriate time. In doing so, "Fred" frees up his human colleagues and reduces unnecessary communication effort within the company.

The digital assistant utilizes Retrieval Augmented Generation (RAG) and generative AI models from the Oracle Cloud Infrastructure. He operates cross-departmentally and thus merges all relevant information – for example from sales, production, logistics and accounting.

"Fred" independently alerts you of important business events and provides tailored recommendations for action. The digital assistant also offers modules for sales optimization, project and personnel resource planning, construction object qualification and much more. “Fred” can process data from all common programs such as Excel or SAP applications. This means that companies can keep their existing IT infrastructure intact and employees can continue to work with their familiar software.

About Brain4Data

Brain4Data GmbH & Co KG in the German town of Saarwellingen develops solutions in the fields of Robotic Process Automation (RPA) and Augmented Intelligence (AI). We analyze communication and processes in companies. On this basis, we identify automation potential that supports a flawless stream of work and communication. We also offer the digital assistant “Fred”, which can be used to automate processes and simplify workflows.

Further information on Brain4Data and the digital assistant "Fred" can be found here:

https://brain4data.de/

Contact

David Woirgardt-Seel

Chief Knowledge Officer

Telefon: +49 (0) 6838 502 09 63

E-Mail: david.seel@brain4data.de