Adesso Intelligent Agent - An Addon Ecosystem for Extending and Empowering Rag Systems

Leon Stolpmann and **Rafael Dubach** and **Stephan Symons** and **Philipp Kuntschik**
leon.stolpmann@adesso.ch

Abstract

Retrieval Augmented Generation (RAG) is a technique that combines natural language interaction with a search and retrieval system connected to a data repository such as enterprise content management systems (CMS) [1]. RAG provides its users with a natural language interface for querying, accessing, and manipulating data contents, such as asking questions, summarizing documents, or classifying text. Leveraging the recent advances in Large Language Models (LLMs) and neural/vector search solutions, RAG for enterprise search has emerged as a key industrial use case. However, RAG also introduces several new challenges and limitations such as verification of truthfulness of given answers (e.g. LLMs tend to "hallucinate" at times [2]), content safety and potential reputational damage caused by harmful answers (in case of public or customer facing systems) and resulting limited trust in these systems. There are also legal and legislative challenges that need to be addressed, as unsafe content [3] or data privacy issues (e.g. PII [4]) are currently still legally ambiguous. To address these challenges, we present an AI framework we call adesso Intelligent Agent that can be adapted to specific industry use cases and provides add-on capabilities to individually address these issues. The adesso Intelligent Agent describes a scalable and customizable RAG-Framework that accelerates development and deployment of functioning systems while allowing to rapidly address current and future challenges such as privacy, compliance, and trustworthiness or the expanding needs from business or environment. Our reference architecture encompasses the ability for deployment on the common cloud providers as well as in a private data center. With this submission, we want to set the focus on our Addon Integration Layer. The layer sits between an outward facing API for Web-interfaces or to be used from other software and the RAG CORE-System and allows for observation, management, and Controlled Intervention of requests to the system. The adesso Addon Library describes a comprehensive set of mechanisms that proved effective working with our clients and allow us to quickly incorporate additional functionality to a RAG system while minimizing the impact on system performance. We will show, how our approach allows to continuously monitor, evaluate, and improve the quality and safety of the generated answers. We believe that RAG has great potential for a wide range of industry specific use cases but also poses significant challenges. By leveraging the best practices contained in the presented framework, companies are able to address these challenges and requirements for compliance in AI systems.