University *of Ljubljana*
Faculty *of Computer and Information Science*

Conversational Agent with Retrieval-Augmented Generation

Katarina Velkov, Nejc Krajšek, Luka Sabotič

Abstract

Keywords
Conversational agent, Retrieval-Augmented Generation

Advisors: Aleš Žagar

Introduction

Large language models (LLM) rely solely on their pre-trained knowledge to generate responses. These models have proved to be powerful but they are prone to generating hallucinated or inaccurate information. Furthermore their use in specialized fields has proved to be challenging as LLMs are trained mostly on publicly available data. That is why the need for more advanced models has emerged. Retrieval-Augmented Generation (RAG) enables LLMs to retrieve more relevant information from various internal databases and web sources during conversations. This process allows for more accurate, domain-specific and up-to-date responses.

The aim of this project is to build a conversational agent using the RAG technique. We will evaluate the performance of our model by comparing it to GPT-4 / BERT models.

Related work

Implementation of Retrieval-Augmented Generation (RAG) was first introduced in 2020 and since then the amount of contributions has been growing. A. Afzal et al. (2024) reported optimization to enhance performance in the academic domain.

Methods

Results

Discussion

Acknowledgments

References