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 wil use an existing large language model, expand it to be able to search the web for relevant topics and use the web data to improve the output. We will evaluate the performance of our model by comparing it to GPT-4 / BERT models.

Related work

The concept of Retrieval-Augmented Generation (RAG) was first introduced in 2020 by [1], who proposed a framework that combines the strengths of large language models (LLMs) with external knowledge retrieval. Since then, RAG has seen significant advancements, particularly in the context of domain-specific applications and conversational agents.

In the paper ”Towards Optimizing a Retrieval Augmented Generation using Large Language Model on Academic Data” by [2], the authors explore various optimization techniques for RAG in the academic domain. They introduce several enhancements, such as Multi-Query, Child-Parent Retriever, Ensemble Retriever, and In-Context Learning, to improve the performance of RAG systems. Their experiments demonstrate that incorporating multi-query techniques significantly boosts retrieval effectiveness, especially in domain-specific contexts like university study programs. The authors also propose a novel evaluation approach, the RAG Confusion Matrix, to assess the effectiveness of different RAG configurations.

Another relevant work is the survey by [3], titled ”Retrieval-Augmented Generation for Large Language Models: A Survey,” which provides a comprehensive overview of RAG’s evolution. The survey categorizes RAG into three paradigms: Naive RAG, Advanced RAG, and Modular RAG. It highlights the importance of retrieval, generation, and augmentation techniques in enhancing the performance of RAG systems. The survey also discusses the challenges and future directions of RAG, such as improving robustness, handling long contexts, and integrating multimodal data.

These works provide a solid foundation for understanding the current state of RAG and its applications in conversational agents. They also offer insights into potential optimization techniques and evaluation methods that can be applied to our project.

Methods

For our project, we aim to develop a conversational agent that leverages RAG to provide accurate and up-to-date responses by dynamically retrieving information from external sources. The proposed methodology involves the following steps:

Dataset Selection

We will use publicly available datasets for training and evaluation. Potential datasets include:

- **MS MARCO**: A large-scale dataset for machine reading comprehension and question answering, which is suitable for training retrieval-based models.

- **Natural Questions (NQ)**: A dataset that contains real user queries and corresponding answers from Wikipedia, making it ideal for evaluating the factual accuracy of our conversational agent.
- **HotpotQA**: A dataset that requires multi-hop reasoning, which will help us test the agent's ability to integrate information from multiple sources.

Implementation Ideas

- **Multi-Query Retrieval**: Inspired by [2], we can implement a multi-query approach to generate multiple variations of the user's query, which will help retrieve a broader range of relevant documents.
- **Ensemble Retriever**: We can combine different retrieval methods (e.g., BM25 and semantic search) to improve the relevance of retrieved documents.
- **In-Context Learning**: We will experiment with in-context learning techniques to fine-tune the conversational agent's responses based on the retrieved context.
- **Prompt Engineering**: We will explore various prompt engineering strategies to guide the model in generating more accurate and coherent responses.

Results

(To be completed after implementation and evaluation.)

Discussion

(To be completed after implementation and evaluation.)

Results

Discussion

Acknowledgments

References

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[2] Anum Afzal, Juraj Vladika, Gentrit Fazlija, Andrei Staradubets, and Florian Matthes. Towards optimizing a retrieval augmented generation using large language model on academic data. *arXiv preprint arXiv:2411.08438*, 2024.

[3] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.