FRI Natural language processing course 2025
Project report

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

Conversational Agent with Retrieval-Augmented Generation

Blaž Grilj, Ana Poklukar, Kristjan Sever

Abstract

The abstract goes here. The abstract goes here.

Keywords
NLP, LLM, RAG

Advisors: Aleš Žagar

Introduction

The development of large language models (LLMs) has significantly advanced the capabilities of conversational agents. However, traditional LLM-based chatbots rely solely on pretrained knowledge, which limits their ability to provide accurate and up-to-date information, especially in dynamic or domain-specific contexts. Retrieval-Augmented Generation (RAG) offers a promising solution by combining real-time information retrieval with generative language models, allowing agents to ground their responses in relevant external data.

In this project, we're building a conversational agent that uses Retrieval-Augmented Generation (RAG) to pull in information from the web in real time. The focus is on answering general knowledge questions—anything from historical facts to current events—with responses that are accurate, coherent, and relevant. Instead of relying only on what it learned during training, the agent will actively search for up-to-date information and use it to generate better answers.

Related Work

Retrieval-Augmented Generation has emerged as a promising approach to enhance the performance of large language models by addressing key limitations such as factual hallucinations, outdated knowledge, and the lack of domain-specific expertise. While LLMs like ChatGPT (OpenAI, 2022) have demonstrated impressive general capabilities (Bang et al. [1]), they are still prone to producing incorrect or outdated information (Cao et al. [2]). These shortcomings are particularly problematic in scenarios requiring factual accuracy and current information.

To address these issues, RAG integrates external information sources—often retrieved via search engines or document databases—into the response generation process. By grounding responses in real-time data, RAG significantly improves the factual consistency and relevance of generated outputs (Guu et al. [3]). This method is particularly effective when paired with robust retrieval systems, which can provide up-to-date context from the vast and constantly evolving content available online.

The field of RAG is evolving rapidly. Gao et al. [4] present a comprehensive survey that categorizes over 100 RAG-related studies into three core research paradigms and outlines the technical challenges across the stages of retrieval, generation, and augmentation. Their work offers a detailed roadmap of the current landscape and future directions of RAG in the context of LLMs.

Complementing this, Chen et al. [5] propose the Retrieval-Augmented Generation Benchmark (RGB), the first benchmark designed to systematically evaluate four core capabilities of RAG-enabled LLMs in both English and Chinese. Their analysis highlights several current limitations of LLMs in RAG settings, including difficulties in accurate retrieval, context integration, and factually consistent generation. The study also suggests concrete directions for future improvement, making it a key contribution to the evaluation of RAG systems.

Initial Idea

Initial Ideas for RAG-Enhanced LLM System

Our plan is to create a conversational agent using Naive RAG implementation [4] for a general knowledge LLM with Wikipedia as the primary external information source.

Model Foundation

Our work will utilize a Large Language Model (LLM) as the foundation for the RAG system. We have two potential approaches:

- Leverage an existing open-source LLM such as LLAMA or similar models, which would reduce computational requirements but may limit customization
- Train a smaller custom model with faculty's high performance compute.

Given resource constraints, we anticipate using a pre-trained model like LLAMA that can be efficiently fine-tuned or adapted to our specific use case without extensive retraining.

IInformation Retrieval Component

The system will incorporate a web scraping module focused primarily on Wikipedia as an information source due to its breadth of knowledge and structural consistency. This will involve:

- Developing a robust web scraper capable of efficiently retrieving Wikipedia articles
- Implementing rate limiting and caching strategies to respect website policies and reduce redundant requests
- Creating parsers to extract relevant content while preserving semantic structure

Query Processing System

A critical challenge will be developing an effective query processing system that can:

- Extract key information needs from potentially ambiguous user prompts
- Transform these needs into specific search queries optimized for Wikipedia's search system
- Handle disambiguation when multiple possible interpretations exist
- Dynamically refine queries based on initial search results
- Provide context from retrieved documents in a LLM frendly way, so not including too much information and ranking it based on relevance.

Evaluation Framework

To measure the effectiveness of our RAG implementation, we will establish a benchmark comparing:

- Performance of the base LLM without retrieval augmentation
- The same model enhanced with our RAG system
- Metrics will include factual accuracy, hallucination reduction rate, response relevance, and response completeness
- Use external benchmarking and ideas from [5]

This comparative analysis will provide quantitative evidence of how external knowledge retrieval impacts model performance across different query types and knowledge domains.

References

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

[2] Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. Factual error correction for abstractive summarization models, 2021.

[3] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020.

[4] 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.

[5] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation, 2023.