



Conversational Agent with Retrieval-Augmented Generation

Matej Belšak, Gorazd Gorup, Luka Bajić

Abstract

Develop a conversational agent that enhances the quality and accuracy of its responses by dynamically retrieving and integrating relevant external documents from the web. Unlike traditional chatbots that rely solely on pre-trained knowledge, this system will perform real-time information retrieval, ensuring up-to-date answers. Potential applications include customer support, academic research assistance, and general knowledge queries. The project will involve natural language processing (NLP), web scraping, and retrieval-augmented generation (RAG) techniques to optimize answer quality.

Keywords

Conversational agent, Retrieval-Augmented Generation

Advisors: Aleš Žagar

Introduction

While Large Language Models (LLMs) have evolved considerably and now produce convincing replies, they have inherent limitations. They rely on training data consisting of documents from the past and may not possess knowledge of current events and developments. Due to differences and properties of training datasets, they may not contain specific domain knowledge, failing to answer certain prompts or outright hallucinating.

Methods

To solve these issues, Retrieval-Augmented Generation (RAG) is used to provide the missing knowledge to the LLM. RAG employs different techniques to retrieve information from external sources based on user's prompt and through prompt augmentation feed the LLM sufficient information to provide informative and factually correct answer. In the survey [1], multiple approaches to RAG are presented, highlighting three architectures: naive RAG, which analyses the user's prompt, retrieves the required information, and appends it, letting the LLM do the rest; advanced RAG, which employs pre-retrieval and post-retrieval modifications to the prompt to make it more suitable for information retrieval and subsequent interpretation by LLM; lastly, modular RAG combines multiple approaches, using iterative prompt enhancement, ranking, fusion, etc.

To better evaluate RAG performance, [2] presents the CRUD framework, employing metrics such as ROUGE, BLEU, precision and recall. Various operations on text (creative gen-

eration from context, usage of information to answer questions, identification and correction of false information, summarization, ...) are measured separately to give a more detailed overview of the model.

For document summarization, LLMs, statistical models, graph-based models and other approaches are used to extract the most important information from text. [3] present multiple solutions, noting that LLMs, while consuming more resources, tend to be more coherent and precise in their summarization if trained correctly.

Recently, a novel LLM, DeepSeek [4], has been presented. Because of its positive benchmarking results, and efficiency due to small number of parameters, it provides a promising starting point for experimentation with knowledge injection and prompt engineering.

In this paper, we focus on implementing a conversational agent operating on knowledge about different art and media. Specifically, the agent is to suggest and converse about films, music, and other media based on user's prompts and preferences. In our contributions, we:

- Develop a conversational wrapper around an existing pretrained LLM, DeepSeek-R1¹;
- Analyse and test prompt engineering techniques on LLM inputs for our defined use cases, noting the placement of information in the prompt, structuring of the prompt, and wording that produces best results;

¹<https://huggingface.co/deepseek-ai/DeepSeek-R1>

- Implement an advanced and/or modular RAG to transform and enhance user prompts and inject necessary knowledge into the final prompt, using approaches such as summarization, ranking, iterative prompt enhancement, and sentiment analysis, via smaller pretrained LLMs;
- Retrieve data from open databases, such as TheMovie-Database (TMDB) ², MusicBrainz ³, and social media platforms, e.g. Letterboxd ⁴;
- Perform benchmarks of our solution with CRUD framework;
- Compare our solution with advanced commercial LLMs, such as ChatGPT.

Results

Discussion

Acknowledgments

References

- [1] 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.
- [2] Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models, 2024.
- [3] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods, 2025.
- [4] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

²<https://www.themoviedb.org/>

³<https://musicbrainz.org/>

⁴<https://letterboxd.com/>