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

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 analyzes 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 generation 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 the small number of parameters, it provides a promising starting point for experimentation with knowledge injection and prompt engineering.

Topic domain

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 and other related media on the user's prompts and preferences. In our contributions, we:

- Develop a conversational wrapper around an existing pretrained LLM, DeepSeek-R1 [1];
- Analyze and test prompt engineering techniques on LLM inputs for our defined use cases, noting the place-

[1] https://huggingface.co/deepseek-ai/DeepSeek-R1

ment of information in the prompt, structuring of the prompt, and wording that produces best results;

- 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 TheMovieDatabase (TMDB) [2], JustWatch [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.

Approach

To develop a reliable conversational agent, we must address several challenges. Our approach focuses on the following areas:

- Understanding user prompts:
  Natural language queries are often ambiguous or incomplete, making effective information retrieval difficult.
- Consistency:
  Conversational agents must maintain logical coherence and avoid contradictions.

Data scraping

Crucial step in the RAG pipeline is the gathering of data (that the model has not been trained on) from external sources, in our case this is achieved via web scraping. Based on the user prompt, our models can access the following data:

- basic information about a film or a person, obtained from TMDB
- a specified number of film reviews, obtained from social media platform Letterboxd, sorted by popularity
- a list of streaming services a specific film is available on

POS tagging and summarization

For the purpose of identifying film titles or names in user's prompt, we use a Roberta-like spacy model en_core_web_trf [5]. We use the same model to summarize retrieved data for advanced RAG.

[2] https://www.themoviedb.org/
[3] https://www.justwatch.com/
[4] https://letterboxd.com/
[5] https://huggingface.co/spacy/en_core_web_trf

Results

In this section we compare the performance of a reasoning DeepSeek-R1-Distill-Llama-8B [4] model and a non-reasoning Qwen3-8B [5] model. For each of these we run experiments on three different sets of parameters: without RAG, with naive RAG - data is scraped and appended to the context, and advanced RAG - data is additionally processed and summarized before being inputed into the model. In total, we are evaluating the following models:

- DeepSeek-R1-Distill-Llama-8B (baseline)
- DeepSeek-R1-Distill-Llama-8B with naive RAG
- DeepSeek-R1-Distill-Llama-8B with advanced RAG
- Qwen3-8B (baseline)
- Qwen3-8B with naive RAG
- Qwen3-8B with advanced RAG

Since we are working with relatively open-ended questions, there is a lack of objective ground truth to evaluate against. Therefore, we construct a set of 50 domain-relevant questions and use manually checked answers from the commercial ChatGPT model as ground truth, so we can apply standard metrics such as ROUGE and BLEU [6]. To avoid overfitting a model to a specific set of questions, we included multiple types of questions into our test set, including yes/no questions, fact checking, list retrieval and summarization. Results are shown in Table 1.

Model	ROUGE	BLEU
deepseek-baseline	0.171003	0.049453
deepseek-naive	0.165164	0.043081
deepseek-advanced	0.16714	0.01866
qwen-baseline	0.192379	0.031395
qwen-naive	0.194950	0.034025
qwen-advanced	0.17380	0.02584

Table 1. Performance comparison against ChatGPT-generated ground truth.

LLM-based evaluation

Metrics such as BLEU and ROUGE are ill-adjusted to our test set, therefore we employ the DeepEval framework [6] in order to obtain more relevant metrics based on context, query and response. Results are shown in Table 2. Metrics related to context are not available for baseline models, because they do not receive any context.

[6] https://huggingface.co/docs/evaluate/index

model	correctness	clarity	answer relevancy	faithfulness	contextual precision	contextual recall	contextual relevancy
deepseek-baseline	0.13	0.05	0.13	0.05	/	/	/
deepseek-naive	0.13	0.05	0.13	0.05	0.13	0.05	0.05
deepseek-advanced	0.13	**0.13**	0.13	**0.13**	0.13	**0.13**	0.05
qwen-baseline	0.13	0.05	0.13	0.05	/	/	/
qwen-naive	0.13	0.05	0.13	0.05	0.13	0.05	0.05
qwen-advanced	**0.13**	0.05	**0.13**	0.05	**0.13**	0.05	**0.13**

Table 2. Performance comparison as evaluated by a 14B parameter Qwen model with GEval framework.

Conversational chatbot

Extending a working model into a conversational chatbot is straightforward: the data retrieval pipeline remains the same, we only need to implement a mechanism for memorization of previous questions and responses. Examples of Qwen chatbot results are shown in Appendix 1.

Discussion

We conclude that it is possible to improve the responses of 8B parameter large language models on queries from a specific domain by employing on-the-fly information retrieval via web scraping.

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.

[5] Qwen Team. Qwen3, April 2025.

[6] Jeffrey Ip and Kritin Vongthongsri. deepeval, 2025.

Appendix

DeepSeek

```
User: How many films did Chuck Jones make?

Qwen: Based on the data provided:

        **Total Number of Films:**
        Chuck Jones directed **over 100 films**, spanning multiple genres and formats.

        ### Breakdown:
        - **Feature-Length Films**:
                - *The Bugs Bunny/Road Runner Movie* (1979)
            - *The Phantom Tollbooth* (1970)
            - *The Grinch Stole Christmas!* (1966)

        - **Shorts and Documentaries**:
            - Approximately **100+ short cartoons**
                - Many of these are Looney Tunes/Merry Melodies cartoons.
                - Includes iconic titles such as *What's Opera, Doc?*, *Duck Amuck*, etc.

            - **Documentaries**: Around **dozens**, e.g.,
                    - *The Magical World of Chuck Jones*
                - *Chuck Jones: Memories of Childhood*

        ### Summary:
        Chuck Jones made approximately **more than 100 films**, primarily consisting of **animated short
             cartoons**, a handful of **feature-length compilations**, and several **documentaries**.
```

Figure 1. Conversation with Qwen