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. *Rephrase and Respond* (RaR) [5] prompting refines user queries by expanding, rewording or clarifying intent, ensuring more effective retrieval and response generation.
- Reducing factually inaccurate output:
  One of the primary issues with LLMs is their tendency to generate misleading or incorrect information. To counter this, we apply *Forward Looking Active Retrieval augmented generation* (FLARE) [6]. The method decides when and what to retrieve to improve accuracy and reduce inaccurate output.
- Consistency:
  Conversational agents must maintain logical coherence and avoid contradictions. *Contrastive Chain-of-Thought* (CCoT) [7] prompting helps models identify patterns and avoid mistakes by comparing correct and incorrect examples. This step is crucial for conversational agents.
- Fine-tuning and optimization:
  To maximize efficiency and response quality, we employ *Automatic Prompt Engineer* (APE) [8] analyzes user input, generates candidate instructions and then uses reinforcement learning to choose the optimal prompt.

[2]https://www.themoviedb.org/
[3]https://www.justwatch.com/
[4]https://letterboxd.com/

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.

Results

In this section we compare the performance of a reasoning deepseek-r1 model and a non-reasoning qwen 1.5 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 with 1.5B parameters (baseline)
- deepseek-r1 with 1.5B parameters and naive RAG
- deepseek-r1 with 1.5B parameters and advanced RAG
- qwen 1.5 with 1.8B parameters (baseline)
- qwen 1.5 with 1.8B parameters and naive RAG
- qwen 1.5 with 1.8B parameters and advanced RAG

Since we are working with relatively open-ended questions, there is objective ground truth to evaluate against. Therefore, we construct a set of 51 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]. Results are shown in Table 1.

Discussion

At the moment, there is a visible improvement in the models' performance when using RAG, however the BLEU and ROUGE evaluation metrics do not emphasize this improvement enough, therefore we intend to devise more representative evaluation methods for our specific domain. Other open

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

Model	ROUGE	BLEU
deepseek-r1 baseline	0.13079	0.00567
deepseek-r1 naive	0.14052	0.01068
deepseek-r1 advanced	0.16714	0.01866
qwen 1.5 baseline	0.18991	0.03687
qwen 1.5 naive	0.19899	0.05811
qwen 1.5 advanced	0.17380	0.02584

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

problems are how to determine which title/person a query is referring to in certain ambiguous cases and how to enable the model to keep a memory in the conversational context.

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] Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. Rephrase and respond: Let large language models ask better questions for themselves, 2024.

[6] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation, 2023.

[7] Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. Contrastive chain-of-thought prompting, 2023.

[8] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers, 2023.