Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240575

Using Retrieval-Augmented Generation to Capture Molecularly-Driven Treatment Relationships for Precision Oncology

Kory KREIMEYER[a,1], Jenna V CANZONIERO[a,b], Maria FATTEH[a,b], Valsamo ANAGNOSTOU[a,b] and Taxiarchis BOTSIS[a]
[a] *Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA*
[b] *The Johns Hopkins Molecular Tumor Board, Johns Hopkins School of Medicine, Baltimore, MD, USA*

Abstract. Modern generative artificial intelligence techniques like retrieval-augmented generation (RAG) may be applied in support of precision oncology treatment discussions. Experts routinely review published literature for evidence and recommendations of treatments in a labor-intensive process. A RAG pipeline may help reduce this effort by providing chunks of text from these publications to an off-the-shelf large language model (LLM), allowing it to answer related questions without any fine-tuning. This potential application is demonstrated by retrieving treatment relationships from a trusted data source (OncoKB) and reproducing over 80% of them by asking simple questions to an untrained Llama 2 model with access to relevant abstracts.

Keywords. Retrieval-Augmented Generation, Precision Oncology, Large Language Models

1. Introduction

The rapidly growing field of precision oncology is often dedicated to determining personalized cancer treatment plans tailored to individual patients based on their clinical phenotype and genotype, characterized by molecular profiling [1]. In practice, identifying these treatments relies on a unique combination of expert medical knowledge, data from the patient's entire clinical and genomic history, and recommendations and recent findings recorded in knowledgebases, meta-knowledgebases, and published literature. This last component is time-intensive, even for experts, and there is considerable interest in developing automated knowledge generation approaches with the goal of turning literature into (actionable) knowledge.

The recent surge of generative artificial intelligence has drawn attention to the application of advanced large language models (LLMs) to biomedicine, but few organizations have the resources to train or fine-tune these models for specific tasks. The technique of Retrieval-Augmented Generation (RAG) [2] can represent a middle ground in which an off-the-shelf (open-source or proprietary) LLM is paired with contextual

[1] Corresponding Author: Kory Kreimeyer, Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, United States; E-mail: kkreime1@jhu.edu.

information pulled from a minimally prepared set of documents. The text from the documents is broken into chunks placed in a vector embedding. Then, based on the actual input query, the most closely related chunks are retrieved and passed to the LLM as plain text that provides additional context for the question at hand.

In this paper, we briefly test the ability of RAG pipelines to capture the same treatments identified by a reliable precision oncology data resource using contextual literature abstracts.

2. Methods

OncoKB is a manually curated, high-quality knowledge base containing, among other knowledge/information, a compilation of therapeutic implications for specific genomic alterations and cancer types [3,4]. We obtained therapeutic relationships for a selection of seven variants (chosen in consultation with oncologists at the Johns Hopkins Medicine molecular tumor board) from OncoKB on 20 March, 2024. For each variant, we used only the information in the therapeutic table from the OncoKB website and determined the cancer type(s) and drug(s) associated with each entry. These were used to form the basis of the treatment queries for the LLM, and to judge the correctness of the responses. We also gathered all PMIDs or abstracts hyperlinked in the OncoKB description column. The text of each abstract was obtained and placed into individual files.

The RAG pipeline was constructed in a DataBricks environment, using a small NCasT4_v3 GPU accelerated cluster (4 vCPU, 1 GPU, 28GiB Memory) running version 13.3 of the DataBricks runtime. The pipeline relied on the LangChain framework [5] for both text processing and model execution.

The abstract files were loaded and chunked into individual text strings with a maximum size of 128 tokens. Then, a vector search index was created in DataBricks, which handled the embedding calculations for the chunks. Only the abstracts associated with a single variant were loaded at one time, and the vector search index was reset between variants.

The final chat model was prepared as a LangChain RetrievalQA chain, with four text chunks to be pulled from the index and using a template with placeholders for both the vector search index results and a user-provided question. The template is shown in Figure 1. One other chain was evaluated as well, which used eight text chunks as context instead of four and is referred to as the Extended Context scenario. The two LLMs used were the open-source Llama 2 70B Chat model and the Mixtral-8x7B Instruct model.

The question that was provided during each run was based on the variant of interest and the cancer types taken from the OncoKB data. It followed the format, "What treatments are recommended for a patient with {cancer_type} with the {variant} mutation?"

Suggested treatments were extracted from the JSON-style output of the models and compared against the expected drugs from the OncoKB data. The matching procedure was fairly lenient, to better simulate a human reviewing the output.

Finally, to compare a non-LLM data extraction pipeline, we used PubTator3, a tool by the National Library of Medicine that recognizes several types of entities and their relations in the literature [6]. These annotations are readily available and contain entity relationships that can convey similar information to the OncoKB therapeutic inferences, albeit at a more granular scale. The PubTator3 output for each PMID of interest was downloaded on 1 April, 2024, although the five non-PMID abstracts cited by OncoKB

could not be found in PubTator. All PubTator3 relations from the abstracts for a specific variant were aggregated and assessed to see whether they contained enough information to support each OncoKB therapeutic relationship.

```
You are an assistant for an oncologist. You are answering bioinformatics,
precision medicine, and genomic questions related to cancer treatment.
Format your response as a JSON object with the following keys:
* treatment: str - The name of the treatment
* rationale: str - Your reasoning as to why the treatment is appropriate

Use the following pieces of context to answer the question at the end:
{context}
Question: {question}
Answer:
```

Figure 1. The prompt template for the RAG pipeline.

3. Results

The seven variants chosen for analysis and the number of singular treatment relationships derived from the OncoKB data are shown in Table 1. The BRAF G466R mutation had the largest number of treatments at eight because the OncoKB data named four specific cancer types with two drugs listed for each (Cobimetinib and Trametinib).

Table 1. The variants used for the analyses and the number of total treatments derived from the OncoKB data as well as the number of abstracts cited by OncoKB that could be obtained for the vector embeddings.

Variant	Derived Treatments	Abstracts Obtained
AKT E17K	3	5
BRAF G466R	8	3
ESR1 L469V	2	5
IDH2 R172G	3	3
PALB2 M723fs33	4	4
PIK3CA E545K	4	6
PTEN P248fs	3	4

As summarized in Figure 2, of the 27 total treatments across the variants, the Llama 2 model captured 16 of them, and the Mixtral model 17 using the initial four context chunks. In the Extended Context scenario with eight text chunks, the Llama 2 model greatly improved and captured 22 treatments while the Mixtral model remained the same at 17 treatments.

Both Llama 2 and Mixtral models followed the requested JSON-style output for all questions, although they used multiple JSON objects when listing multiple treatment items. Both frequently produced extra free text after the requested JSON-style format as

well. These sections usually provided disclaimers about the recommendations or sometimes additional rationale for the choice of treatments.

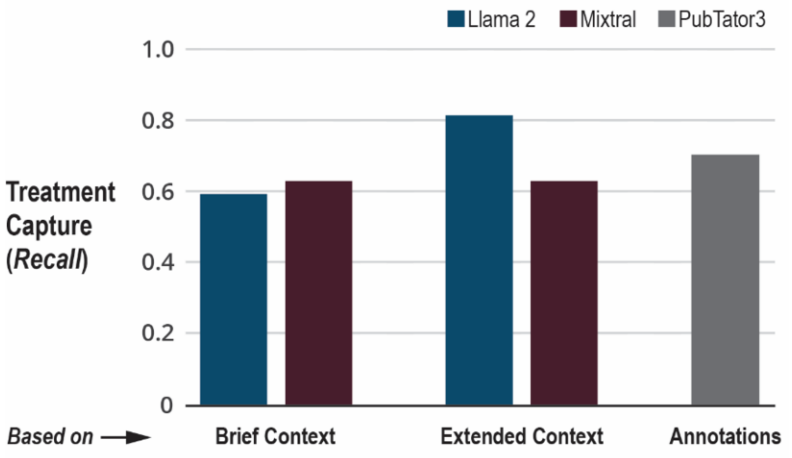

Figure 2. The performance results of the Llama 2 and Mixtral LLMs, as well as PubTator3, at providing treatment recommendations that reproduced the variant-specific OncoKB-derived treatments.

The PubTator3 relations derived from the abstracts were found to support 19 of the OncoKB treatment items. These were usually relations of the form (CHEMICAL, treat, DISEASE) and often supported further by a (GENE, associate, DISEASE) relation and/or a (CHEMICAL, negative_correlate, GENE) relation, indicating an inhibitor. These results had to be interpreted at the gene level because PubTator3 did not identify any relations for the variants of interest except for AKT1 E17K.

4. Discussion

In the best instance, a RAG pipeline with the Llama 2 model reproduced over 80% of the simple OncoKB confirmed therapeutic relationships across seven variants when seeded with relevant literature abstracts. RAG pipelines are simple to create, requiring only a collection of literature or similar text sources. Furthermore, this demonstrates they can be effective even when handling only abstracts, even though the therapeutic implications confirmed by OncoKB may be embedded within the papers. The RAG approach can also outperform the concept level relationship detection of PubTator3 in drawing therapeutic conclusions.

As an exploratory study, our work has a few limitations that point toward avenues for further research. The selection of literature was limited to papers already known to be related to the variant of interest, allowing us to demonstrate the (high) sensitivity of a RAG approach. We believe that other applications could find small and relevant literature data sets as well, though, through well-targeted querying methods. Secondly, the use of abstracts rather than full text papers may limit the data available. OncoKB curators would have reviewed the entire documents when verifying the therapeutic

evidence, but access to full text literature is often highly restricted for data mining uses like these pipelines. Finally, we did not spend extensive time on prompt engineering, but we do not believe this is a vital step for this application, especially since the outputs always followed the requested JSON-style formatting. We did test a shorter variant of the template, but found very few notable differences.

When using eight text chunks of context instead of four, the Llama 2 model netted six additional correct treatments. The larger context had a better chance of pulling in relevant information, as in the case of the PIK3CA E545K material, where it was necessary to even see the name of one of the target drugs (RLY-2608), which had not appeared in the first four chunks. However, the new text may also prove misleading, as in the case of the PALB2 M723fs33 mutation where new statements about Rucaparib may have have caused the model to trim out its previous correct responses of Olaparib and Talazoparib plus Enzalutimide for prostate cancer.

We did not attempt to differentiate between the OncoKB levels of evidence. The models did provide rationales for the treatment recommendations as requested, but parsing these to determine the true level of evidence behind the treatment would likely be subjective. This would be more recommended if searching for newer treatment signals. Many of the variant and therapy relationships we selected have been documented in literature for many years.

5. Conclusion

We believe that RAG methods show promise in reproducing oncology treatment knowledge from OncoKB with a simple pipeline that does not require LLM fine-tuning. Further studies could explore whether providing more timely information could identify emerging (or even novel) treatments. Automatic generation of this kind of knowledge may expedite oncologists' literature review and decision-making steps.

Acknowledgments

This work was supported by the National Cancer Institute (NCI) [U01CA274631].

References

[1] Schwartzberg L, Kim ES, Liu D, Schrag D. Precision Oncology: Who, How, What, When, and When Not? Am Soc Clin Oncol Educ Book. 2017(37):160-9. doi: 10.1200/edbk_174176.

[2] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of the 34th International Conference on Neural Information Processing Systems; Vancouver, BC, Canada: Curran Associates Inc.; 2020. p. Article 793.

[3] Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology. 2017(1):1-16. doi: 10.1200/po.17.00011.

[4] Suehnholz SP, Nissan MH, Zhang H, Kundra R, Nandakumar S, Lu C, et al. Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. Cancer Discovery. 2024;14(1):49-65. doi: 10.1158/2159-8290.Cd-23-0467.

[5] Chase, H. (2022). LangChain [Computer software]. https://github.com/langchain-ai/langchain

[6] Wei C-H, Allot A, Lai P-T, Leaman R, Tian S, Luo L, Jin Q, Wang Z, Chen Q, Lu Z. PubTator 3.0: an AI-powered Literature Resource for Unlocking Biomedical Knowledge. arXiv preprint arXiv:240111048. 2024. doi: 10.48550/arXiv.2401.11048.