

Evaluating the Effectiveness of Open-Source Large Language Models in Qualitative Healthcare Research using RAGAS

Nikhita Bhatt¹, Manon M. Schladen, PhD²

¹Montgomery Blair High School, Silver Spring, MD; ²The Catholic University of America

Abstract

Large language models (LLMs) are showing promise as a tool that can be used for qualitative healthcare research. We analyzed 20 transcripts of interviews with veterans, caregivers, and clinicians regarding the use of robots in veteran caregiving using Meta's Llama 2. We evaluated the effectiveness of the analysis using the RAGAS framework, to see whether LLMs could effectively analyze transcripts. We concluded that Llama 2 has a RAGAS score of 0.69, indicating moderately good performance.

Introduction

Large language models (LLMs) are becoming more widely used in research, showing promise in qualitative research, especially. However, LLMs have also been shown to include hallucinations, or generating irrelevant answers.¹ More research needs to be done to understand whether and how LLMs can be used to accurately and efficiently interpret qualitative data. Retrieval augmented generation (RAG) is a framework that allows users to provide custom context to an LLM, and the LLM uses that context to better frame the answer.² To evaluate the performance of a RAG framework, one can use RAGAS³, which is a tool that can evaluate the effectiveness of an LLM on a variety of metrics, and return quantitative values for each.

Methods

The data utilized for the RAG pipeline included 20 interview transcripts, where questions were designed to determine possible use cases and requirements for robots in veteran healthcare. The AI model that was being tested was Meta's Llama 2. This project compared three retrieval chains (methods that retrieve the most relevant elements of the context to answer a question) to determine which is most effective. The questions were generated by OpenAI's ChatGPT LLM. The chains included basic RetrievalQA chain, a parent document retriever, and an ensemble retriever. The RetrievalQA chain indexes the transcripts of the veteran interviews, analyzes the query, and then processes it. It then ranks the contents of the document by relevance through semantic similarity techniques. Using the passages that it ranks as most relevant, the system creates a final response. The parent document retriever is a type of RetrievalQA chain that focuses more on larger parts of context, and is used more to find relevant information rather than to answer specific questions. The ensemble retriever combines multiple types of algorithms to create a retriever. We measured four metrics (values ranging from 0 to 1) using the RAGAS framework, which we used to compare the efficacy of the three retrieval pipelines. The metrics included: faithfulness (whether an answer is accurate given the context), context precision (whether the information pulled from the context is relevant), context recall (whether all the relevant information was retrieved), and answer relevancy (whether the given answer was relevant to the prompt). An overall RAGAS score was also calculated, which was the harmonic mean of the four metrics.

Results

The retrievalQA chain had a faithfulness, answer relevancy, context precision, and context recall score of 0.7, 0.97, 0.6, and 0.6, respectively. The ensemble retriever had scores for these same metrics of 0.66, 0.78, 0.7, and 0.6. Finally the parent retriever had scores of 0.7, 0.4, 0.7, and 0.8. Table 1 shows overall the RetrievalQA chain performed the best with a RAGAS score of 0.69, and the parent retriever performed the worst with a score of 0.62.

Discussion

We set out to determine whether Llama 2 was effective in transcript analysis and, if so, which retrieval chain was the most effective. All three retrieval chains had moderately good performance (RAGAS scores of > 0.62) meaning Llama 2 showed promise in being effective for qualitative research. However, we expected the ensemble retriever to work more efficiently than the other two retrievers because it uses multiple types of algorithms. However, results show that the RetrievalQA chain is more effective than the ensemble retriever. This could be due to the fact that the disjoint structure of this data could be easier for RetrievalQA to parse. We can conclude that LLMs show promise in analyzing qualitative data, but future research should be directed toward choosing appropriate retrieval pipelines.

Table 1. RAGAS Metrics for Each Retrieval Chain

Retrieval Chain	Faithfulness	Answer Relevancy	Context Precision	Context Recall	Overall
RetrievalQA	0.7000	0.9710	0.6000	0.6000	0.6906
Parent Retriever	0.6967	0.4000	0.8000	0.8000	0.6216
Ensemble Retriever	0.6662	0.7766	0.7000	0.6000	0.6798

Figure 1. Retrieval Chain Flow Diagram

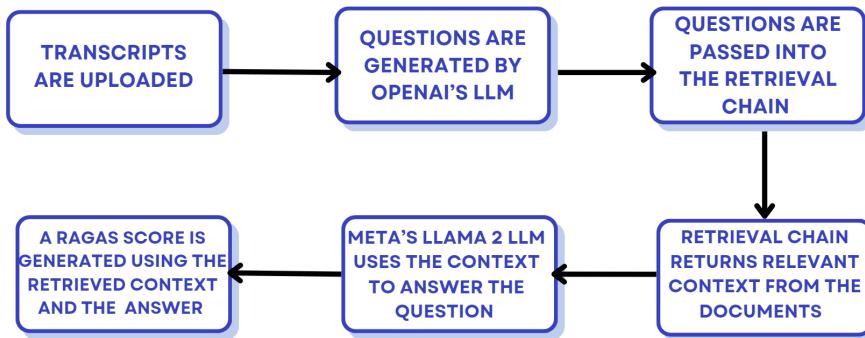


Figure 2. Example of a Question, Context, and Answer from RetrievalQA

Generated Question	Generated Answer	Retrieved Context (From Interview Transcript)
What challenges do you face in providing healthcare services to veterans?	Some of the challenges we face in providing healthcare services to veterans include repetitive interactions due to various issues, lack of staffing, gaps in service provision between different locations, especially for walk-in traffic, and difficulties in navigating the VA system. Additionally, caregivers may face challenges in helping veterans get from one location to another. It can be a daunting task for new veterans to understand and access care at the VA, which may prevent them from continuing their care. The VA is working on improving in-house care and reducing costs for care in the community to make it a more attractive option for veterans compared to the private sector.	<p><i>Interviewee:</i> "Repetitive interactions that you have with veterans and there's lots of different issues. You know, just the business of clinics on any given day, there's a lot of gaps in service provision from one location to another, especially for walk-ins. For any kind of tasks that you know would be you know anything from helping somebody fill out advanced directives to how do I apply for this benefit."</p> <p><i>Interviewer:</i> "So what are the challenges that caregivers for your patients experience?"</p> <p><i>Interviewee:</i> "Trying to help the veterans get from one location to the next. [The VA is] a big machine and it takes a while to figure that out. I'm also keenly aware that that is a daunting task for new veterans... to the VA that sometimes precludes them from continuing their care at the VA. We are continuing to push, you know, more in house care, trying to reduce care in the community costs. I'm trying to bring in more veterans, and be a more attractive option than the private sector to veteran care."</p>

References

1. Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods*, 23. <https://doi.org/10.1177/16094069241231168>
2. Lewis, P. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Arxiv*. <https://doi.org/https://doi.org/10.48550/arXiv.2005.11401>
3. Es, S., James, L., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. *Arxiv*. <https://doi.org/https://doi.org/10.48550/arXiv.2309.15217>