Retrieval Augmented Generation: Leveraging Large Language Models and External Knowledge for Natural Language Search

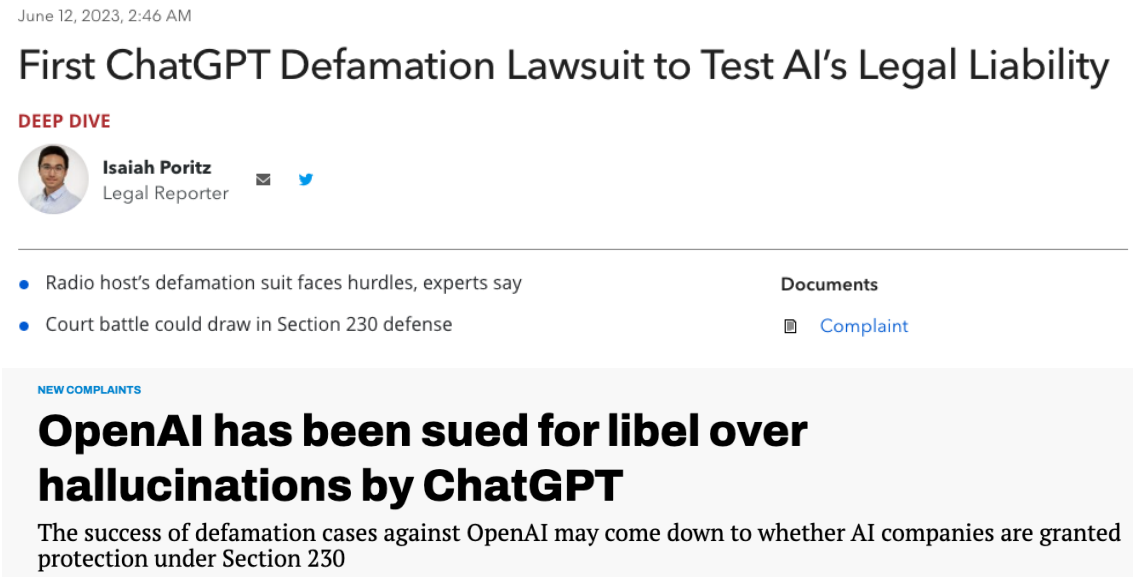
RAVINDRA SADAPHULE, Johns Hopkins University, USA

Traditional information retrieval produce answers to user's natural language query by searching through corpus of documents or images crawled and indexed ahead of time. It uses technologies like inverted index, Tf-IDF and okapi bm25 to search through millions of documents and produce relevance results in a sub second. The results are produced as set of documents or images with highly relevant documents at the top. The user need to read title, teaser text to see what's relevant to their queries and click on relevant documents to get more information.

There are few known challenges that are inherent in traditional information retrieval system. The first challenges is that the retrieval is done using Bag of Words method. Given a set of n words in the query, the system looks for all permutations and combinations of the words in the documents and surfaces the results. For example, the system cannot differentiate between queries like "chocolate milk" and "milk chocolate". The second challenge is that the answer that user is looking for could be scattered through multiple documents. This requires user to make additional effort to comb through the documents, aggregate the response from various documents in the search result to generate the result.

With the recent advances in large language models, we could address the challenges described above by combining large language models and information retrieval engine to generate a unique one answer that is enough for user to satisfy his/her intent. However if we just large language models out of the box, it won't yield us relevant and up to date results. Traditional seq2seq models often generate generic, templated responses due to the lack of external world knowledge and reasoning capabilities. These models have been trained on public corpus. They are not aware of unique enterprise knowledge. As a result Large language models suffers from hallucinations and their knowledge base has a cut off date (Chat GPT has cut off date Sep 2021). As a result they can not be used to search for real time information that enterprises typical has.

Consequence of hallucinations have been so severe that few companies have been sued. Here are few examples.



June 12, 2023, 2:46 AM

First ChatGPT Defamation Lawsuit to Test AI's Legal Liability

DEEP DIVE

Isaiah Poritz
Legal Reporter

- Radio host's defamation suit faces hurdles, experts say
- Court battle could draw in Section 230 defense

Documents

Complaint

NEW COMPLAINTS

OpenAI has been sued for libel over hallucinations by ChatGPT

The success of defamation cases against OpenAI may come down to whether AI companies are granted protection under Section 230

The primary reason behind these hallucination is that LLMs are not grounded. Traditional large language models like GPT-4 embeds general world knowledge base and natural language understanding together as single module. The training of these models is every expensive in terms of time and cost. GPT-3 model training on A100 GPU cost about $100M and one month. These large language models often have a cut off date. (e.g. GPT-3 has cut of date of Sep 21, 2022). Hence they are unable to comprehend latest changes in the world events.

In this paper we show that by combining information retrieval and large language models together. We can address limitations of large language models and traditional information retrieval system and generate a single answer that's highly relevant to user's query. This is still very nascent technology known as Retrieval Augmented Generation (RAG). RAG models work by first retrieving relevant context documents from an external knowledge source like Wikipedia or enterprise corpus in response to the input text. These retrieved documents provide key facts, concepts, entities, and other knowledge that can inform and ground the text generation process. The retrieved documents are encoded and provided as additional input to the seq2seq model which then uses them to generate an output response.

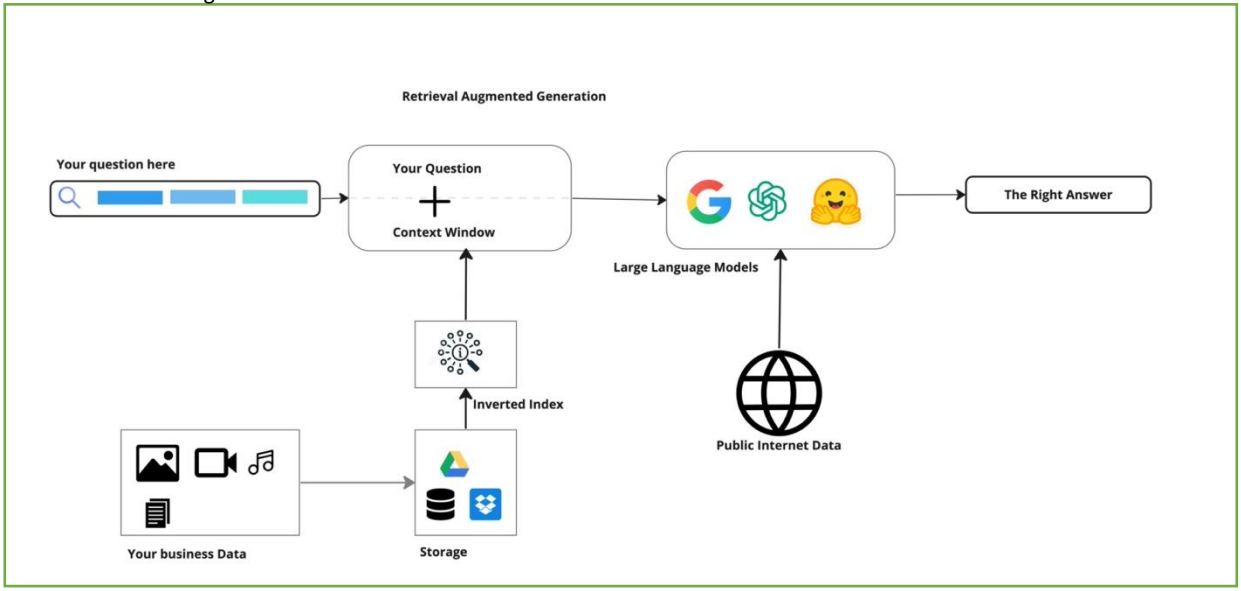
Proposed research will evaluate different encoder architectures and retrieval mechanisms to identify optimal techniques for retrieving salient, relevant external knowledge and integrating it into the text generation process. The goal is to develop RAG models capable of producing more factual, knowledgeable, and coherent natural language compared to traditional seq2seq models. We'll compare results with and without RAG and show unique strengths about RAG approach.

Additional Key Words and Phrases: Large Language Models, Retrieval Augmented Generation, Semantic Search, Pretrained Models

1 INTRODUCTION

As hype on Generative AI continues, just using Gen AI technologies out of the box to solve information retrieval challenges has not yielded meaningful results. There are challenges in terms of LLM hallucination and high training cost and limited context window. In Additional as prompt engineering evolves, users are looking for answers for long queries and Traditional IR engines are not great at finding answers for long questions as they use typical BM25 search which is based on BOW (bag of words model) [Ref]. Traditional Search also generate 20 blue like that user has to comb through to get to result. We could generate a unique answer that summarizes text from 20 blue links thus improving productivity of the user.

There is unique opportunity to combine power of semantic search and large language model to solve informational retrieval challenges. Here is how it in intended to work.



A typical enterprise has his data stored in variety of formats and in multiple stores that could include cloud storage and on promise storage. In order to make the data searchable, the first thing that we need to do is to ingest the data into an inverted index like Solr or Elastic Search. The data will ingested as plain text and embeddings will be generated on the fly and will be stored in the inverted index along with metadata. When user types in a query, first we'll generate an embedding for the query and search the inverted index using embedding search method. We'll get semantically related documented as search result which are much more relevant than traditional bm25 results. We'll retrieve textual metadata from these embeddings. We'll then collate metadata from first 10 documents of search result and create context for LLM prompt. We'll add question as prompt and search results as context window and then issue the query to large language model. This context window helps ground the LLM. We'll ask LLM to generate final answer only in the context of current window. LLMs will help us generate one answer in a way that's more intuitive appealing to the user. LLMs can also augment the data from public internet but in the context of search results that we retieved.

1 IMPLEMENTATION

- I'll be using OpenAI's API to query GPT -3.5 LLM embeddings API for inference. I have OpenAI subscriptions. I have checked rate limiting by issuing calls using Open AI tokens from the python notebook. I can get upto 100 requests per minutes and 50,000 tokens per minute. This is good enough for my experiment
- I'll be using Elastic Search open source version 8.10 as inverted index. I'll be downloading it locally on my machine and create single node single shard index
- I'll be indexing 1,000- 10,000 documents from specific public domain (travel, insurance, wikipedia) and Indexing them in elastic search
- I'll create Python Streamlit application to showcase how Retrieval Augmented Search works end to end.

REFERENCES

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[2] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268.

[3] Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases?. arXiv preprint arXiv:1909.01066.

[4] Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., ... & Dolan, B. (2018). QuAC: Question answering in context. arXiv preprint arXiv:1808.07036.

[5] OpenAI. 2023. GPT-4 Technical Report. arXiv:cs.CL/2303.08774

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.