Currently generative AI is undoubtedly one of the hottest topics in the tech field, while large language models (LLMs) are models pre-trained on massive, diverse data, mostly to generate natural language, i.e. human-like text. ChatGPT is an exceptional example of LLMs that has outstanding ability to generate answers to users' queries. Those LLMs are widely used among individual users and businesses. For real-world application, especially under business scenarios, it is usually necessary to ground the content generated by an LLM on some specific knowledge domain. One of the solutions is to first extract contextual information from domain-specific knowledge bases and transform the query by attaching the extracted context to it and then input the transformed query to LLMs. For my capstone project, I will work on context retrieval, an important component of this solution. Accordingly, this literature review focuses on two fundamental, influential papers in this area.

The first is *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* by Meta AI researchers. In this paper, Lewis et al., (2020) proposed retrieval-augmented generation (RAG), an architecture for question-answering language models. Specifically, the architecture consists of a retriever and a generator. The retriever retrieves the most relevant documents from extra knowledge bases given an input (usually a query from a user), while the generator generates the final output (an answer to the user's query) after taking in the retrieved documents and the original input. In their experiments, they broke the articles in Wikipedia December 2018 dump into 100-word chunks as an extra knowledge base and vectorized those chunks using BERT. On top of that, they used several popular question-answering datasets for training, validation and test and also used BERT for inputs (questions) encoding. With those data preprocessed, a pre-trained bi-encoder was set up as the retriever to retrieve the top-k chunks, and BART-large was set up as the generator to generate the output based on the retrieved chunks and input. They then adopted an end-to-end method to jointly train the retriever and the generator and fine-tuned the input encoder and the generator. Finally, the results of their experiments showed that "RAG models generate more specific, diverse and factual language than" (Lewis et al., 2020) a single generator.

The retriever used in Lewis et al., (2020)'s experiments was based on DPR. DPR, Dense Passage Retriever was proposed by Karpukhin et al., (2020) in the paper *Dense Passage Retrieval for Open-Domain Question Answering*. As a breakthrough from the traditional information retrieval models like TF-IDF and BM25 for context retrieval, Karpukhin et al., (2020) proposed to represent candidate contexts and inputs (queries) in dense vectors and proved that powerful dense representations could be learned through training. In their experiments, they kept only unstructured data (text) from the Wikipedia December 2018 dump and split the articles into 100-word chunks as candidate contexts and used five question answering datasets for training. They used two BERT networks as encoders, which were respectively used to convert candidate contexts and inputs into dense representations. The idea of the training process was, with good encoders, positive(relevant) contexts should be more similar to a given input than negative(irrelevant) contexts, and they measured similarity by inner product. According to the results of their experiments, DPR outperformed the benchmark substantially in terms of top-20 accuracy.

Those two papers are of great significance for the application of LLMs. Though they were initially for open-domain question answering, the architecture and methods are widely used to ground LLMs on domain-specific knowledge or any information of interest. In both papers, the key to context retrieval is getting dense embeddings of contexts and queries through training, which could be time and computing resources consuming. Since these two papers were published, the generative AI landscape has been progressing significantly with LLMs getting much more powerful. Meanwhile, a domain-specific knowledge base may not be as large as a Wikipedia dump. Therefore, simple methods like the classic information retrieval models might be able to perform well, in combination with the currently cutting-edge LLMs. This is the area that I am going to explore in this project.

References:

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.
https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 6769-6781). (EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference). Association for Computational Linguistics (ACL). https://aclanthology.org/2020.emnlp-main.550.pdf