Argonne National Laboratory is a scientific research institution owned by the U.S. Department of Energy. Argonne's operations teams must manage a massive corpus of documents, which creates a massive bottleneck on the efficiency of the Lab, prompting interest in more effective knowledge management and retrieval systems.

To address these challenges our group has worked on refining a Retrieval Augmented Generation (RAG) pipeline that is capable of retrieving text relevant to a user query to allow an LLM to respond with factual, Argonne-specific knowledge.

This quarter, we have made progress on a few key components of the pipeline. Firstly, we have built a working prototype of a computer vision pipeline for segmenting documents by sections. In addition, we have constructed a prototype of a RAG pipeline that is capable of generating hypothetical documents, which can help us more directly locate relevant document segments with our current measures. We also experimented with the use of text summarization to make retrieved documents more palatable for the end user. Finally, we built a pipeline that uses an LLM as a judge to automate the process of labeling document segments as semantically relevant and compared these results to our cosine similarity measures.
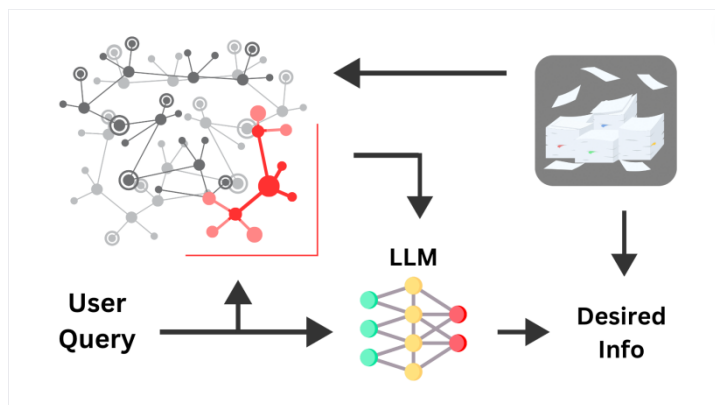
Fig. 1 Pipeline Overview Diagram
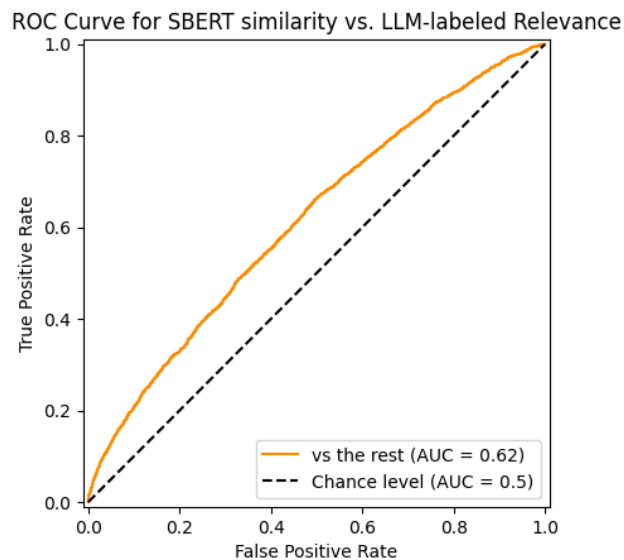
Fig. 2 ROC Curve for SBERT Similarity and LLM Relevance