Implementing a Large Language Model Interface to the IFPRI Technical Report Library

Eric M. Furst and James Michaels

The IFPRI report library spans nearly 40 years of technical documents. The current report interface implemented through the IFPRI website is capable searching for indexed terms, authors, and project identifiers, which is stored mainly in the form of webpage metadata. With the recent growth of large language models (LLMs), new methods of processing and generating text in natural language are rapidly emerging and present an opportunity to build a more capable interface to the IFPRI library. We propose that IFPRI initiate a half-year test project to prototype an LLM-based application built to query the IFPRI report library. Our goal is to enable deeper semantic search across the library, making the content of reports more accessible and, in the future, possibly open it to data mining and generative synthesis.

The project will implement retrieval augmented generation (RAG), illustrated in Figure 1. RAG stores chunks of documents in a vector database and uses the similarity of these to vectorized queries from a user. Top matches between the vector database and query are used as input to a general LLM, which generates a natural language response to the user. For instance, the response could be in the form of a summary of matching or relevant documents and the associated metadata, such as links to the online IFPRI reports.

Through its own outcomes and reports back to the membership, the project will also provide the opportunity for IFPRI community to learn about the capabilities and limitations of these AI tools in technical work.

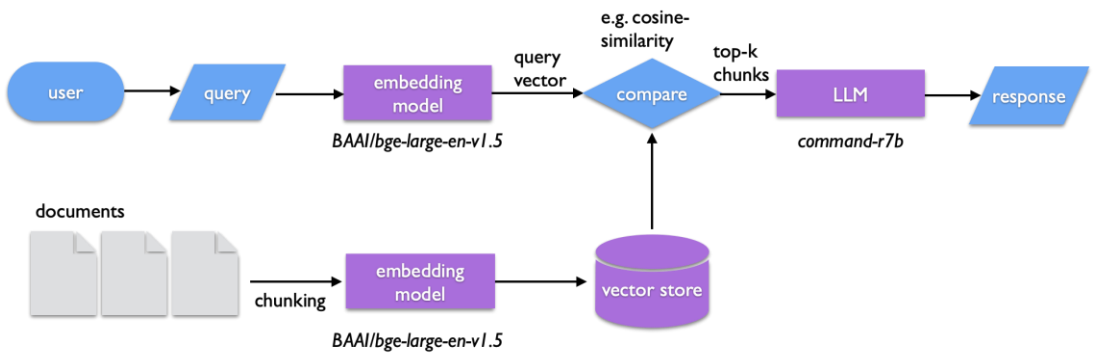

Figure 1. Retrieval augmented generation uses a vector store (or vector database) of embedded chunks of documents. The similarity between the stored vectors and an embedded user query returns information to a language model that processes the natural language response to the user.

Project deliverables

Furst has experience with building RAG applications using local models. He will advise a data science master's student to implement the RAG query. Principal aims and milestones include:

1. Build a vector database of a subset of the IFPRI technical reports, which will initially focus on machine-readable file formats. (Older reports are stored as images and require further processing.)
2. Implement query processing, including query rewriting strategies.
3. Build the LLM response workflow that generates responses to a user query.
4. Test and evaluate the performance of the RAG query and search and make recommendations for further work, including interactive modes (a chatbot), protection against malicious prompts and jailbreaking, and the development of a web-accessible search application.

To protect IFPRI intellectual property, we plan to use locally run language models for vectorization, query processing, and generation. All data will be stored locally and will not use remote data services.

Project cost and duration

$10,000 funds are requested to support one data science master's student for six months.