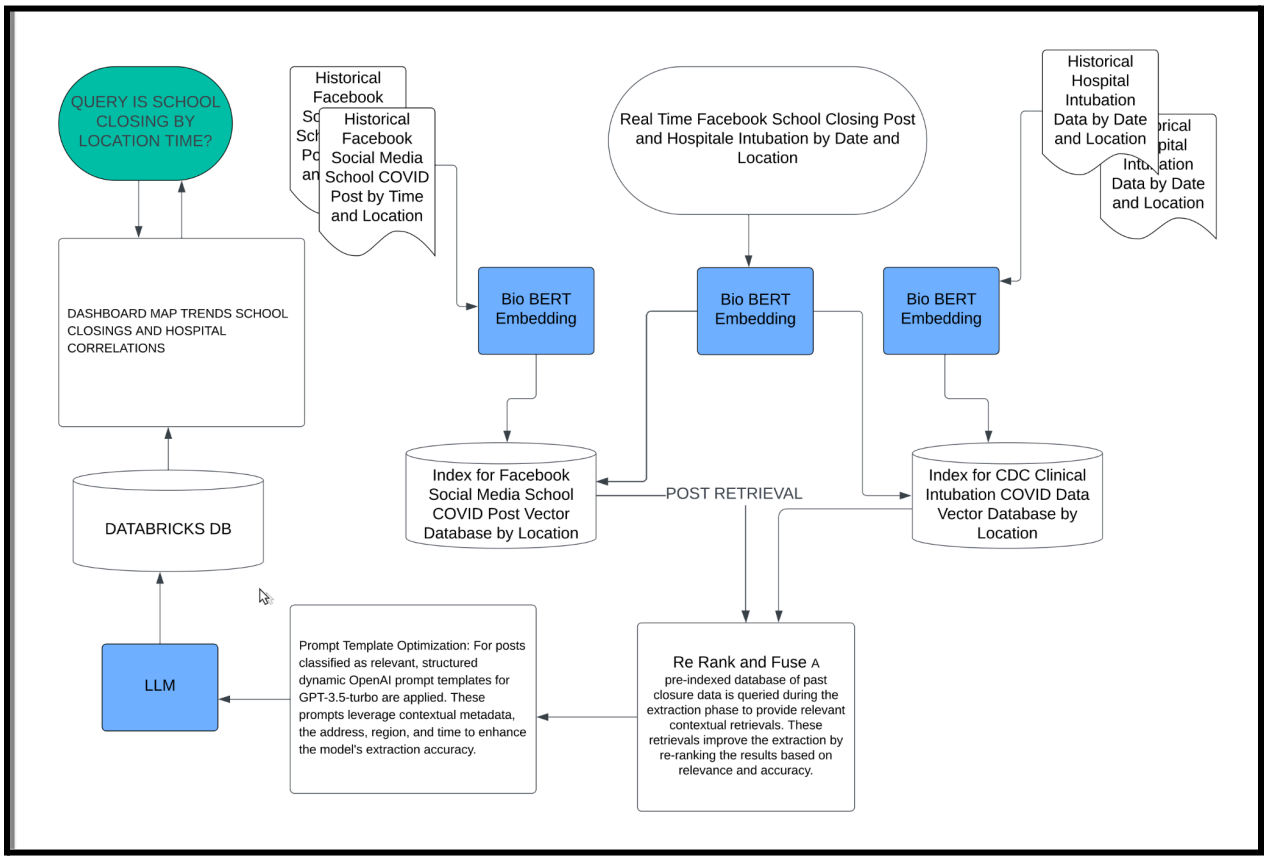
Center of Disease Control - LLM RAG Architecture to Predict School Closures due to COVID

https://github.com/tmlrnc/LLM/blob/main/FacebookLLMDatabricksCDC



My Schoolclosure LLM Pipeline automates the tracking of school closures across the U.S. by leveraging Databricks, Machine Learning (ML), and Generative AI tools. The system processes Facebook posts from schools and integrates clinical data of COVID intubation cases, using an LLM-based architecture for data analysis and report generation. The architecture also incorporates OpenAI prompt optimization and Retrieval-Augmented Generation (RAG) re-ranking for improved accuracy and relevance.

LLM Architecture:

1. Data Ingestion and Preprocessing:

- Data Sources: The pipeline ingests Facebook posts using the Facebook Graph API, capturing unstructured text about potential school closures. Additionally, clinical data for COVID intubation cases is ingested from hospital databases via REST API calls.
- Preprocessing: Irrelevant posts are filtered using regular expressions. Clinical data is anonymized and formatted for integration with the school closure data. Text inputs are preprocessed for tokenization to prepare them for the language models.

2. OpenAI Prompt Optimization and RAG Re-Ranking:

- BERT Model for Classification: A fine-tuned BERT-based model classifies posts as relevant or irrelevant to school closures.
- Prompt Template Optimization: For posts classified as relevant, structured dynamic OpenAI prompt templates for GPT-3.5-turbo are applied. These prompts leverage contextual metadata, the address, region, and time to enhance the model's extraction accuracy.
- RAG Re-Ranking: A pre-indexed database of past closure data is queried during the extraction phase to provide relevant contextual retrievals. These retrievals improve the extraction by re-ranking the results based on relevance and accuracy.

3. LLM-Based Analysis:

- Extraction Phase: Posts classified as relevant are analyzed by the GPT-3.5-turbo model via the Azure OpenAI API, extracting structured closure-related information such as closure dates and causes. The extracted information is combined with relevant retrievals from the RAG process to ensure contextual accuracy.
- Post-Analysis: The extracted closure information is enriched with metadata such as geographical locations and timestamps from the schools or districts.

4. Integration of Clinical Data:

- Data Enrichment: The pipeline integrates clinical data (e.g., COVID intubation cases) with school closure data based on shared attributes like region and date. This allows for correlation analysis between school closures and spikes in intubation cases.

- LLM-Based Correlation: The combined dataset is analyzed using the GPT model to identify potential correlations between school closures and COVID trends.

5. Reporting and Storage:

- Storage: The final results, enriched with metadata and clinical data, are stored in a Databricks database.
- Dashboarding: Reports are generated daily or weekly through Databricks SQL queries, allowing for easy visualization of trends and patterns related to school closures and clinical outcomes.
- Reporting on Trends: By combining school closure data with clinical information, the pipeline generates reports that highlight patterns and trends, such as correlations between spikes in intubation rates and school closures.

ML Pipeline Flow Summary:

1. Ingest Phase: Facebook posts and clinical data are ingested and stored.
2. Filtering Phase: Irrelevant posts are removed.
3. Analysis Phase: Posts are classified and relevant data is extracted using LLMs with optimized prompts.
4. Extraction Phase: Relevant closure information is standardized and refined through RAG re-ranking.
5. Enrichment and Reporting: Final enriched data is stored and visualized for reporting.

By incorporating OpenAI prompt template optimization and RAG re-ranking, the pipeline ensures high accuracy and contextually relevant outputs, allowing stakeholders to monitor and report on school closures effectively. The extension to integrate clinical data enables further insights into how health trends may influence school closure decisions.