Chandler Davis

Github Repository: <https://github.com/ChandlerD1025/8-K-Document-LLM-Analysis>

LLM Document Analysis Report

I approached this assignment by first accessing SEC’s API to map stock tickers to their corresponding Central Index Key (CIK). Using this mapping, I pulled the most recent 8-K filings for each company. The raw HTML documents were parsed with Python’s BeautifulSoup library, and I extracted the first 5,000 characters of text from each filing to optimize for LLM processing. Any filings that failed to load or had invalid formatting were skipped.

To increase the likelihood of detecting product announcements, I selected 19 companies known for frequent innovation and retrieved their 8 most recent 8-K filings. Each document was then passed to a locally hosted LLM using the Ollama API. After testing DeepSeep-r1:8b on llama2, and Mistral 7B, I found that the **Mistral 7B** model produced the best results. My prompt instructed the model to return either a product name and a short description (180 characters or fewer), or null for both if no product was found. Consistent formatting, prompt design, and pre-processing were essential to maintaining accuracy.

Mistral was able to successfully extract new product data from several companies. For each filing, a row was written to a CSV containing the company’s CIK, ticker symbol, filing time, new product name, and a product description. Examples of extracted products include Apple’s iPhone 16, Netflix’s upcoming shows, the NVIDIA Hopper H200, and Google’s Wiz security platform. Out of 150 filings analyzed, 41 contained product or service announcements. Given that 8-Ks often focus on financial or executive updates, I found these results to be optimal.

During the development process, I encountered several challenges related to both SEC data access and LLM prompting. To retrieve filings from the SEC website without being blocked, I had to add a custom header containing a user-agent and email address; using a placeholder email worked for this purpose. The larger challenge, however, was getting the LLM to return only the relevant information. Specifically, a new product name and a short description, without any financial information or commentary. This required multiple prompt changes and testing with different models. Ultimately, the issue was mostly resolved by creating a highly specific prompt for the Mistral model that clearly instructed it to ignore all non-product-related information and to format the response consistently.