**Mia Bruno**

**3/26/2025**

**LLM Document Analysis**

**CAP5619**

**Intro**

In this report, I will explain how I extracted findings from an automated examination of SEC 8-K filings that detect new product releases and announcements from publicly traded companies. This program makes the process automatic by combining web scraping, text extraction, and LLMs. Its setup is meant to be effective, dependable, and quickly expanded.

**Methodology**

My methodology involved a few pivotal steps, starting with gathering company ticker symbols and CIK numbers from the SEC's website that was provided. After extracting the tickers and CIK numbers, I grabbed 8-K filings for a selection of companies by implementing the CIK numbers to find the correct documents that would show the information needed. I then dug through these filings, focusing on sections "Item 7.01" and "Item 8.01" because they usually have new product announcements. An LLM processed the extracted text to pinpoint and retrieve new product names and descriptions from the data. The data on extracted products was organized into a structured format and saved as a CSV file to facilitate analysis and reporting.

**Data Preprocessing**

The data preprocessing phase included the execution of several essential steps. The CIKs were zero-padded until all achieved a consistent 10-digit length, which met the SEC filing URL format requirements. The HTML/XML structure of SEC filings received parsing treatment from BeautifulSoup, which stripped out unnecessary tags and formatting elements to generate an optimized input for the LLM. The research applied regular expressions to extract "Item 7.01" and "Item 8.01" sections from filings because these sections had the greatest likelihood of including new product release information.

**Model Usage**

At the center of extracting product details was the deepseek-r1:1.5b LLM, which I accessed via the Ollama library. By giving the LLM precise instructions to analyze the filing content, I was able to pinpoint the name and a short description of any new products. The code specified that it was analyzing an SEC 8-K filing, included the company's identification details, and required the extraction of both the "New Product Name" and "Product Description." It also specified that a response of "No new product found" should be given if no new product information was present, limiting the product description length to 180 characters. The LLM's response was then processed using regular expressions to pull out the "New Product Name" and "Product Description."

**Results**

Data for analyzing the first 100 company listings came from the SEC's company tickers JSON file. The script found and collected information about [Number of Products Found] new product filings. The extracted product data was stored in a CSV file named sec\_8k\_product\_releases.csv. The CSV file consists of key details including the company name, stock ticker, filing date, and the names and descriptions of each identified new product if there was any.

**Challenges**

I faced several challenges while I was creating and executing my code. The SEC rate-limiting policy created difficulties since it restricted data retrieval volumes during specific periods and retry logic with exponential backoff became essential to prevent SEC server blocks. Building a flexible extraction system encountered obstacles because SEC filings contain inconsistent formatting and structural variations. Developers must meticulously design parsing logic and regular expressions to manage data variation effectively. Continuous improvement is necessary for the extract\_section function since its data extraction depends on precise regular expression patterns. The primary obstacle when using LLMs involved developing precise and understandable prompts since their performance relied heavily on these elements. LLM outputs often contain hallucinations and mistakes that need comprehensive verification and filtering. In conclusion, communicating with an LLM to analyze multiple SEC filings requires significant computational power and extensive memory resources making this project extremely challenging for someone with minimal code experience.