Christopher Robles

Spring 2025

AI for FinTech

LLM Document Analysis

<https://github.com/ChristopherRobles/LLM-Document-Analysis>

This project aims to extract information about new product releases from SEC 10-K filings using large language models (LLMs) and structure the extracted data into a tabular format saved as a CSV file. A 10-K filing is an annual report submitted by publicly traded companies to the U.S. Securities and Exchange Commission (SEC) that provides a comprehensive summary of their financial performance, including details on new products, risks, and operations. Due to better data availability, I opted to use 10-K filings, which allowed for the extraction of the most relevant information. The extracted data includes the company name, stock ticker, filing timestamp, product name, and product description.

The 10-K filings were obtained from the SEC’s EDGAR database. These filings contain comprehensive financial reports, which also include mentions of new product releases. The data collection process involved accessing and downloading 10-K filings in text format from the SEC EDGAR website using provided URLs. The CIK numbers were retrieved from the SEC’s publicly available JSON file, and Python’s BeautifulSoup was used to scrape additional company information.

Preprocessing involved removing irrelevant sections of the 10-K filings to focus on product mentions. The regex library was used to help identify and extract relevant information from sections that typically contained product-related details. Once the relevant text was isolated, it was processed using a locally installed LLM, accessed through the Ollama library, the specific model used was deepseek-r1:1.5b. The LLM was prompted to identify sections mentioning new products and generate descriptions, limited to 180 characters. Python-based queries were used to interact with the LLM, and the structured data was saved in CSV format.

The CSV file generated included columns such as the company name, stock ticker, filing timestamp, new product name, and a short description of the product. A total of 104 rows of data was generated from 17 companies, with each row representing a product release. Due to hardware and model constraints, I had to run the CIK numbers individually, as running several at once would either take too long or cause the program to crash. The focus was on S&P 500 companies to extract relevant product information.

Several challenges were faced during the project. Initially, I planned to use Named Entity Recognition (NER) with Python’s SpaCy library to automatically detect company and product names. However, I encountered difficulties training the NER model for this task and ultimately removed this step. As a result, I relied solely on section-based text extraction using the LLM. Incorporating NER in the future could improve precision in identifying relevant entities. Additionally, the model response times were slower than anticipated, and getting a concise product description proved challenging. From the 17 companies analyzed, the CSV output showed that only a handful of companies had usable product descriptions.

Initially, I planned to use 8-K filings, but due to difficulties in accessing them in bulk, I switched to 10-K filings. This shift required adjusting the focus of the extraction since 10-K filings cover broader financial details rather than exclusively product announcements.

Despite these challenges, the code was able to extract product release information from SEC 10-K filings, structure the data into a CSV format, and generate the required outputs. Future improvements could involve processing a larger number of filings, improving the model's accuracy in detecting more nuanced product mentions, and optimizing the speed of LLM queries, potentially by using faster local models or cloud-based APIs.