SEC 8-K New Product Extraction-Methodology Report

Aasiyah Alli

March 26, 2025

Project: LLM-Based SEC 8-k filing analyzer

Objective:

The goal of this project is to analyze SEC form 8-k filings to automatically identify and extract new product announcements using large language models (LLMS). Each extracted instance is structured into a tabular format for further analysis.

Methodology:

1. Ticker to CIK mapping

The model uses the SEC- provided JSON file (https://www.sec.gov/files/company\_tickers.json) to convert company tickers (e.g., AAPL) into CIK numbers required to query filings. This mapping is cashed locally to avoid repeated calls.

1. Querying 8-k Filings

Using the CIK, I retrieve up to 100 recent 8-k filings for each company via the SEC’s Atom feed. Each feed entry contains the URL and is used to extract full filing text.

1. Filing Text Extraction

Each 8-k HTML document is parsed with BeautifulSoup to find the correct document link (usually .htm or .html). The full page is then converted to plain text using .get\_text()

1. New Product Extraction via LLM

The model utilizes the OpenAI GPT-3.5 Turbo model to extract:

* company name
* new product
* product description

1. NER considerations

Although the prompt handles most extraction, the model also loads spaCy’s en\_core\_web\_sm model to help with Named Entity Recognition (NER) for company and product names. However, the final version of the pipeline uses the LLM exclusively for extraction.

1. Scalability and Rate Limiting

To avoid hitting SEC rate limits or API throttling, we introduced a 1-second pause between document requests.

1. Technologies used:

* Python 3
* equests, BeautifulSoup4, csv, json, re, os, spacy, openai, logging
* OpenAI GPT-3.5 Turbo API
* spaCy NER model

1. Scale and Scope

* Tickers: 3 (AAPL, GOOGL, MSFT)
* Documents parsed: Up to 100 per ticker

1. Challenges

* Some filings did not contain product announcements
* inconsistent formatting of SEC filings led to parsing failures