**Scrape the 1975 Pacific hurricane season Wikipedia page**

Methodology for 1975 Pacific Hurricane Season Data Extraction: The approach started by scraping the 1975 Pacific hurricane season Wikipedia page using the **BeautifulSoup** library in Python. The table containing storm data was extracted directly from the page, and relevant columns, including storm names, dates active, areas affected, and deaths, were parsed.

To structure the scraped data, I leveraged an LLM (**Llama 3**) to format the data into a structured table with the necessary columns: storm name, date start, date end, areas affected, and deaths. The LLM processed the raw data and provided it in a markdown table format. Afterward, the table was converted to CSV format and the headers were adjusted for clarity. The final CSV file contains only the required data points for easy use. In terms of data quality, care was taken to ensure accurate table extraction using **Regex** and proper column alignment. The LLM was guided to ensure no extraneous data was included. Any potential data discrepancies (such as missing values for deaths) were left as they appeared in the source to preserve the integrity of the dataset.

I used API from this source: <https://replicate.com/account/api-tokens>

Website scraped: <https://en.wikipedia.org/wiki/1975_Atlantic_hurricane_season>

**Usage**

* First create your API key (takes 30 seconds, register via GitHub)
* Download the libraries I used: **pip install -r requirements.txt**

Date: 17/09/24

Author: Ilias Adamidis