Methodology

# Data Scraping Approach:

The process began by scraping data from the 1975 Pacific hurricane season Wikipedia page. Using requests, I retrieved the raw HTML content of the page, which was then parsed using BeautifulSoup. My focus was on extracting the relevant information for each storm, including the storm names, dates, affected areas, and the number of deaths.  
  
For each storm, I extracted the heading (storm name) and all related paragraphs containing its details. These paragraphs were stored as a list associated with the storm name. To clean the data, I removed any unwanted elements, such as citation references (e.g., "[1]"), and handled non-breaking spaces and other encoding issues using regular expressions. This ensured that the scraped text was clean and structured before further processing.  
  
The scraped data was stored in a JSON file, where each entry contained the storm name and its associated content. This JSON file served as an intermediary step, allowing for easier processing and extraction using an LLM in the following phase.

# Data Extraction and Structuring:

For data extraction, I utilized OpenAI's API to transform the unstructured content into structured data. The JSON file created from the scraping process was used as input to the LLM. The model was instructed to extract specific fields: hurricane/storm name, start date, end date, areas affected, and number of deaths. This ensured that relevant information was extracted, even when it was embedded in descriptive paragraphs.  
  
I also implemented logic to handle missing data. For example, if no deaths were mentioned, the default value of '0' was used. If no areas were explicitly listed as affected, I returned 'not known' instead of leaving it blank. The extracted data was then formatted consistently and stored in a CSV file for easy access and analysis.

# Data Quality Assessment:

To ensure the quality of the extracted data, several steps were followed:  
- Completeness: I ensured that all required fields (storm name, start date, end date, areas affected, and deaths) were present for each hurricane entry. Missing fields were handled with default values, such as 'not known' for affected areas and '0' for deaths.  
- Consistency: I standardized the date format using a custom date formatting function, ensuring that all dates followed the same dd/mm/yyyy structure. This provided consistency across the dataset.  
- Cross-validation: After extraction, I cross-checked the data against the original text on the Wikipedia page to verify accuracy. This included ensuring that the areas affected, dates, and deaths were correctly interpreted and extracted by the model. Any discrepancies were corrected during this review process.  
  
The combination of structured data extraction and thorough data quality checks ensured that the resulting dataset was both accurate and reliable for analysis.