Progress report

Project scope update

At the beginning of the project, I planned to analyze user activity data for ChatGPT and DeepSeek from several Kaggle datasets. However, after reviewing the datasets more carefully, I realized that the user activity values were produced by predictive models rather than real measurements. For example, the dataset included DeepSeek usage data before the product was even released in early 2025, which made the data unreliable for analysis.

Because of this issue, I updated my project scope and shifted to analyzing AI-related news instead. The new direction focuses on collecting real-time news articles for topics such as “ChatGPT,” “DeepSeek,” and “Artificial Intelligence” using API-based data sources. This ensures the data is authentic and suitable for trend analysis.

Data source

The primary data source for the revised project is the GDELT 2.0 Document API, which provides real-time global news coverage in a structured JSON format. Using targeted keyword queries such as “ChatGPT,” “DeepSeek,” and “Artificial Intelligence,” the API returns article-level metadata including titles, URLs, and publication timestamps. Although the API’s schema varies slightly depending on the specific query and the availability of information, it consistently supplies enough core metadata to support trend analysis and exploratory evaluation of AI-related news activity.

Issue / difficulties

During the early stages of data collection, I encountered several challenges when attempting to gather AI-related news directly from various websites. Many sources implemented aggressive anti-scraping measures, such as dynamic content loading, rate limiting, and bot-detection mechanisms, which made consistent extraction nearly impossible. In addition, several websites did not provide accessible or stable endpoints for retrieving structured article data, resulting in incomplete or unusable outputs.

Due to these limitations, I shifted to using the GDELT 2.0 Document API, which offers standardized access to large volumes of global news content without the restrictions commonly found on individual websites. This significantly improved the feasibility and reliability of the data collection process.