

## **Section 1: Scraping Data**

Step 1: Extract links and metadata from the Excel file



Step 2: Scrape PDF text corpus, exclude irrelevant cases



## **Section 2: Definition Extraction**

Step 3: Use an LLM to extract the market definition section



Step 4: Use LLM to isolate individual market definitions



## **Section 3: Cleaning & Aggregation**

Step 5: Clean JSON outputs and merge into final dataset