My Role and Responsibilities in this Web Crawler

General

- Scrape through website(s); might be through a *.txt list or a single URL

Technical

- Scrape, adhering to the following *parameters*:
  - `DEPTH`: Decides the recursive crawling depth
    * A webpage has `n` hyperlinks; each `i`th hyperlink has `k` hyperlinks within
    * `DEPTH = 2` scrapes `n * k` links
    * `DEPTH = 3` scrapes `n * k * j` (exponential growth; computationally expensive)
    * **Design Conundrum**: Should `DEPTH` be seed-specific or seed-agnostic?
  - `MAX_PAGES`: Global limit; stops after `n MAX_PAGES` pages scraped
  - `BASE_CONCURRENCY`: Number of concurrent browser tabs (e.g., via Selenium/Playwright)
  - `DELAY_JITTER_MIN` / `DELAY_JITTER_MAX`:
    Jitter mimics human behavior to avoid detection and rate-limiting.
    Defined as:

$$jitter(p) = \sum_{p=1}^{n} \frac{t_p - t_{avg}}{n}$$

    These parameters set a random delay range `[MIN, MAX]` between requests.
    Purpose:

    1. Avoid detection
    2. Mimic human behavior

3. Prevent rate-limiting

- `SEED_PAUSE_xxxxx`: Delay between crawling different seed URLs (mimics human behavior)