# Approach Taken

## Preparation & Setup

• Implemented an EnhancedFinploySitemapGenerator class for modularity and maintainability.
• Added a session with custom headers to mimic a browser and avoid bot detection.
• Initialized sets for discovered, crawled, and failed URLs to prevent duplicates.

## Seed URL Collection

• Checked for existing sitemaps via /sitemap.xml, /robots.txt, and related paths.
• Parsed sitemaps recursively to gather initial URLs.
• Added base URLs to the crawl queue.

## Smart URL Guessing

• Generated potential URLs using job-related keywords, locations, and category patterns.
• Validated guessed URLs with a fast HEAD request before queueing.

## Crawling & URL Extraction

• Used BFS-style crawling.
• Extracted links from <a> tags, data attributes (data-url, data-href, etc.), and JavaScript content using regex.
• Applied strict validation to filter irrelevant or harmful links.

## Output & Reporting

• XML sitemap – structured for SEO submission.
• JSON report – includes categories, success rate, and elapsed time.
• Log file – detailed crawling history.