AI Specialist & Data Engineer (Researcher position)

Technical Review and Analysis of Task Requirements

Skills Required:
To complete this assessment, you'll need to demonstrate skills in the following areas:

1. **Web Scraping** - Extracting raw HTML data from various sources.
2. **Data Classification** - Using AI tools to extract and structure information.
3. **Data Modeling & Storage** - Designing schemas and selecting the right databases.
4. **Architecture & Design Patterns** - Understanding high-level system design.

Pre-Execution

Before executing your solution, please provide a high-level explanation covering the following:

1. Architecture

Describe your approach to system design, including:

- Key components (e.g. schedulers, servers, AI tools, storage).
- Flow structure (web scraping → data classification → storage).
- Whether you're considering serverless architecture, message queues, pub/sub systems, or other methods to optimize for scalability and efficiency.

2. Web Scraping

Share your approach to scraping, including:

- Tools you're familiar with (e.g. Playwright, Puppeteer, BeautifulSoup).
- How you would handle dynamic HTML content (e.g. via XPath, rendering engines).
- Known limitations in scraping and how you would overcome them (e.g. rate limits, captchas, IP blocking).

3. Data Classification

Detail your strategy, including:

- Tools you prefer to use (e.g. GPT, spaCy, HuggingFace) and why.
- Your approach to prompt engineering - how do you manage context, reduce cost, and maintain efficiency when dealing with large amounts of raw data?

4. Data Modeling & Storage

Explain:

- What type of database you would choose and why (e.g., SQL vs. NoSQL)
- What data model fits the task best and how you would structure the information.

Post-Execution

Once your solution is ready, you should:

- Be able to explain it clearly from end to end.
- Demonstrate a simple "happy flow" to show how your system works.
- Show awareness of potential scalability challenges and outline ideas for long-term improvements.
- **(Bonus)** Use tools like **Lucidchart** to document your architecture or flow visually.