AI Engineer – Take Home

Task:

Build a generalized directory scraper that can handle multi-page directories and extract structured data. This task involves heavy data scraping, combined with LLMs, and other techniques for normalizing, cleaning, and enriching the raw information.

Here are some example directories we will test against:

- https://sais.health.pa.gov/commonpoc/content/publicweb/nhinformation2.asp?COUNTY=Allegheny
- https://sdpsych.org/Find-a-Psychologist
- https://profiles.stanford.edu/browse/school-of-engineering?p=1&ps=100
- https://psychologyhouston.org/directory.php
- https://community.bapapsych.org/search/newsearch.asp?bst=&cdlGroupID=&txt_country=&txt_statelist=&txt_state=&ERR_LS_20250827_222102_27698=txt_state%7CLocation%7C20%7C0%7C%7C0
- https://math.berkeley.edu/people/graduate-students
- https://www.ycombinator.com/companies/

The scraper should be flexible enough to work across different directory formats.

Input / Output Specification

You will receive a structured dictionary as input that describes the fields we want to extract. Example:

```
{
  "name": "name of the student",
  "title": "example: Postdoctoral Scholar, Bioengineering",
  "email": "their email address",
  "page_url": "url of their page, e.g. https://profiles.stanford.edu/jijumon-a-s",
  "bio": "their bio"
}
```

When run on Stanford's Engineering directory, your scraper should return all 6,297 profiles in a clean, structured list of dictionaries that match the input format.

Notes

- The scraper should be generalized and not tailored only to the provided examples. We will also test it on other examples.
- Performance matters. It should be reasonably fast.
- You are free to use LLMs, DOM parsing, or other creative approaches.
- Browser agents such as browserbase or browseruse are allowed but can be slow or unstable. Feel free to use it in combination with other techniques.
- 100% accuracy is not required, but the more directories it can handle effectively, the better.

Be prepared to discuss design decisions and tradeoffs. If you had more time, what would you improve?

Next Step

Take a sample of 100 rows from the scraped PhD student dataset and pass them through the **Sixtyfour /enrich-lead** endpoint. Enrich on a few interesting fields and prepare to share feedback and improvement ideas.

When you log in to app.sixtyfour.ai, you will automatically receive $5 in API credits. If you need more, let us know.