




Python Job Scraper Requirements




1 The script must:

-  **Show the correct progress counter**
 - It should display per-page progress (e.g., "10/30 kept=0 skip=10"), *not* cumulative totals.
 - `processed` should only count *real job detail pages*, not skipped or invalid ones.
 - Each Glassdoor page should display exactly 30 results (or the number of results available on that page).
-  **Always display a job title** in the console/log output
 - Even if the page HTML doesn't contain a title, it should fall back to something human-readable (like parsing the last slug of the URL).
 - Titles should appear in all `KEEP` and `SKIP` log messages.
 - There should be no blank `Title` entries anywhere.
-  **Show Title, Company, short description on KEEP**

The `log_keep` implementation already prints:


 - Line 1: `[KEEP] {Title} at {Company} [badge]`
 - Line 2: `{short description snippet}`
 - Line 3: job or apply URL

Wire the three inputs:

- `title`: your parsed title or slug fallback
- `company`: your parsed company
- `description_snippet`: first ~120 characters from the job text
-  **Stop printing excessive or irrelevant SKIPs**
 - Don't show lines for "URL skipped by rule" (listing/search URLs that aren't real jobs).
 - Only log true job decisions (KEEP or SKIP after analysis).
-  **Fix the `urlparse` crash**
 - The 'cannot access local variable 'urlparse' error must be resolved by removing the local import statement inside `main()`.
-  **Auto-backup the Python file**
 - Creates a timestamped copy of **every** `.py` file in your project/Code Archive folders.
 - Skips backup entirely if you are running a script from inside the `Code Archive` folder.
 - Keeps **all backups indefinitely** until you decide to clean them up later.

Keep only the global import:

```
from urllib.parse import urlparse, urljoin, parse_qs, unquote
```

-  **Ensure the script still logs progress even on errors**
 - The progress bar should continue updating after each link, even if an exception is thrown for one job.

Code Integration Requests

You asked for:

- Clear, step-by-step instructions for *where* to add or move code.
- Exact line placement guidance (e.g., right after `listing_links = find_job_links(html, page)`, or right above `return {...}` in `extract_job_details`).




- Confirmation of whether large code blocks (highlighted in screenshots) should be deleted or modified.
- Assurance that edits don't break other parts of the scraper (you didn't want to rewrite major logic from scratch).
- Example of what the final combined code block should look like.

Functional Expectations






- The program must still:
 - Scrape Glassdoor and similar sites usually.
 - Distinguish between "Product Owner / Product Manager" and irrelevant titles.
 - Collect and store job details (Title, Company, Location, Posted, etc.) correctly.
 - Skip duplicates and previously visited jobs.
- You still want to export all results to your Google Sheet and save them as CSV files afterward.

Code Quality / Behavior Requests

You wanted:

-  **Consistency in logging** — keep `[INFO]`, `[SKIP]`, `[PROGRESS]`, etc. standardized.
-  **Clear and minimal console output** — no extra diagnostic clutter (only meaningful skips and keeps).
-  **No regression of working features** — the title extraction fix and progress fix shouldn't break the rest of the script.

What You *Didn't* Want

-  No "quick patches" that make the code look messier or introduce new side effects.
-  No changes that affect your exports (Google Sheets / CSV) or your `results` data structure.
-  No new global counters or external files to fix progress — it must be handled within `main()`.
-  No removing your existing detailed logging system (you still want to see reasons, counts, etc.).
-  No complete rewrite of `extract_job_details()` — only fixes and fallbacks for Title.

Summary of the Concrete Fixes Requested

| Area | Change Needed | Description |
|---------------------|--|---|
| <code>main()</code> | Fix <code>urlparse</code> crash | Remove local import, keep top import |
| <code>main()</code> | Add per-page counters | Introduce <code>kept0</code> / <code>skip0</code> and adjust <code>progress()</code> call |
| <code>main()</code> | Only increment <code>processed</code> for valid jobs | Move <code>processed += 1</code> after skip filters |
| <code>main()</code> | Silence URL-rule skips | Don't print/log <code>should_skip_url(link)</code> messages |
| Logging | Always show a title | Add <code>title_for_log()</code> helper and use it for KEEP/SKIP |

```
extract_job_details(  
)
```

Guarantee a fallback
Title

Add slug-based fallback before each
`return {...}`

Output

Keep progress alive

Update `finally:` block to call
`progress()` reliably
