**1. Prompting Strategy & Pipeline Structure**

**Prompting Strategy:**

* **Material Classification:** The LLM is guided to identify materials using:
  + Explicit schema definition in Pydantic model
  + Context injection through field descriptions
  + Predefined materials list in the system prompt
  + Constrained output format ("must be one of the predefined materials")
* **Handling Ambiguity:** For material matching:
  + LLM performs semantic matching (e.g., "AA batteries" → "Household Batteries")
  + Partial matches resolved through category context
  + Unmatchable materials are excluded per instructions

Key components:

**Pipeline Structure**

1. **Playwright:** Handles dynamic content rendering
2. **HTML Sanitization:** Removes scripts/styles to reduce noise
3. **Extraction Chain:** create\_extraction\_chain\_pydantic enforces output structure
4. **Post-processing:** Maps materials to categories using predefined lookup

**2. Edge Case Handling**

|  |  |
| --- | --- |
| **Edge Case** | **Solution** |
| **Nested HTML** | Semantic parsing by LLM ignores structural complexity |
| **Map-Only Data** | Playwright ensures full DOM rendering before extraction |
| **Inconsistent Labeling** | LLM's semantic understanding handles variations (e.g., "Cellphones" → "Cell Phones") |
| **Missing Fields** | Required fields in Pydantic model enforce completeness |
| **Dynamic Content** | Explicit wait for .result-item selector |
| **Material Ambiguity** | Multi-step matching: material → category → standardized label |

