This is a test file

SYSTEM\_PROMPT\_MERGE = """

You merge two extracts that describe the same PDF page:

(1) vis\_text – produced by Vision OCR, rich in tables/figures

(2) raw\_text – produced by a PDF text extractor, reliable for plain text

Return strict JSON with keys:

  merged\_text  : best reading order text

  omissions    : list of fragments present in only one source

  table\_json   : list of tables; each table is a list of rows (each row is list of cells)

Rules:

- Prefer vis\_text for tables and images.

- Prefer raw\_text for headers, footers, running paragraphs when duplicates occur.

- Do NOT add or hallucinate content.

"""

"""

mixed\_layout\_extractor.py

Author: Jun Xu

Robust PDF extractor that routes pages with large figures or tables

through GPT-4o Vision and merges the result with classical text extraction.

"""



|  |  |  |  |
| --- | --- | --- | --- |
| item | categories | | comments |
|  | A | B |  |
| 1 | a.1 | b.1 | Test result1 |
| 2 | a.2 | b.2 | Test result 2 |
| 3 | a.3 | b.3 | Test result 4 |



