

PDF table extraction test

A text-layer document for checking row order, column labels, wrapped cells, page continuation, and safe handling of merged cells.

1. Standard rows and columns

Document	Pages	Review priority	Expected output
Quarterly report	24	High	Four aligned Markdown columns
Research paper	12	Medium	Every row remains separate
Product brief	6	Low	Header appears once
Invoice batch	18	High	Values stay under the right label

Expected check

The generated Markdown should contain four columns and four complete body rows. No value should move into a neighboring column.

2. Wrapped and multiline cells

A visual line break inside a cell should remain part of that cell rather than becoming a new table row.

Element	Source text	Review note
Long description	A wrapped sentence that occupies multiple visual lines but belongs to one logical cell.	Keep as one cell
Manual line break	First line Second line	Use a safe in-cell break
Pipe character	Budget forecast	Escape the pipe in Markdown
Empty value		Keep the empty column position

Expected check

Wrapped text stays in the Source text column. The manual line break may be represented with an HTML break inside a Markdown table cell.

3. Cross-page continuation

This table continues onto the next page. The repeated header is intentional and should not become a second unrelated table.

Item	Owner	Status	Evidence
Row 01	Editorial	Review	Source note 01
Row 02	QA	Open	Source note 02
Row 03	Research	Ready	Source note 03
Row 04	Editorial	Review	Source note 04
Row 05	QA	Open	Source note 05
Row 06	Research	Ready	Source note 06
Row 07	Editorial	Review	Source note 07
Row 08	QA	Open	Source note 08
Row 09	Research	Ready	Source note 09
Row 10	Editorial	Review	Source note 10
Row 11	QA	Open	Source note 11
Row 12	Research	Ready	Source note 12
Row 13	Editorial	Review	Source note 13
Row 14	QA	Open	Source note 14
Row 15	Research	Ready	Source note 15
Row 16	Editorial	Review	Source note 16
Row 17	QA	Open	Source note 17
Row 18	Research	Ready	Source note 18
Row 19	Editorial	Review	Source note 19
Row 20	QA	Open	Source note 20
Row 21	Research	Ready	Source note 21
Row 22	Editorial	Review	Source note 22
Row 23	QA	Open	Source note 23

Item	Owner	Status	Evidence
Row 24	Research	Ready	Source note 24

4. Merged-cell safe fallback

Markdown pipe tables do not express row spans or column spans. This section checks that a complex visual header is flattened without silently inventing structure.

Market	Quarterly metrics	
	Revenue	Growth
North America	\$1.2M	18%
Europe	\$0.9M	13%
Asia Pacific	\$0.7M	21%

Expected check

A converter should preserve the visible values and warn that merged cells were flattened. The source PDF remains the reference for the original spanning layout.