

Converting a PDF into Markdown sounds like a formatting chore, but the hard part is really reading order. A two-column report interleaves its text streams, so a naive extractor will splice the left column's third paragraph into the middle of the right column's heading, and the output reads like a corrupted transcript rather than a document.

Tables fail differently. The visual grid you see on the page is only a hint; the PDF stores text runs and line strokes, and the converter must reconstruct which cells belong to which row before it can even think about pipes and dashes. When header cells span two columns, most tools emit a flattened row that loses the grouping entirely.

Scanned pages are the third failure mode. There is no text to extract, only an image, so the pipeline has to hand the page to an OCR or vision model and hope the handwriting, stamps, and skewed margins come back as coherent sentences instead of confident nonsense scattered across the middle of an otherwise clean document.

Quarterly invoice summary

Item	Qty	Unit Price	Total
Consulting hours	12	85.00	1020.00
Data export fee	3	40.00	120.00
Template license	1	190.00	190.00
Support retainer	6	55.00	330.00

This page exists only as a picture. A scanner produced it, the PDF embeds it as a bitmap, and no amount of text extraction will recover these words without OCR.