Table 3 Overview of text and table extraction tools used in our study. Key extraction capabilities include extraction of Image (I), Text (T), Metadata (M), Table of Contents (TOC), and Table (TB). Most tools use rule-based (RB) technology, with some offering Optical Character Recognition (OCR) capabilities. However, Nougat and Table Transformers were not the primary focus of this study.

Tool	Version	Extraction	Technology	Output
PyPDF	4.3.0	I, T, M	RB	TXT
pdfminer.six	20240706	I, T, TOC	RB	TXT, HTML, hORC, JPG
PyMuPDF	1.24.7	I, T, TB	RB(MuPDF), OCR	TXT, HTML, SVG, JSON
pdfplumber	0.11.2	I, T, TB	RB(pdfminer)	TXT, HTML, hORC, JPG
pypdfium2	4.30.0	T	RB	TXT
Unstructured	0.14.10	T, TB	RB, OCR	TXT
Tabula	2.9.3	TB	RB	DataFrame, CSV, JSON
Camelot	0.11.0	TB	RB	DataFrame, CSV, JSON, HTML
Nougat	base(350M)	T	Transformer	Markdown
Table Transformer	TATR-v1.1-All	TB	Transformer	Image

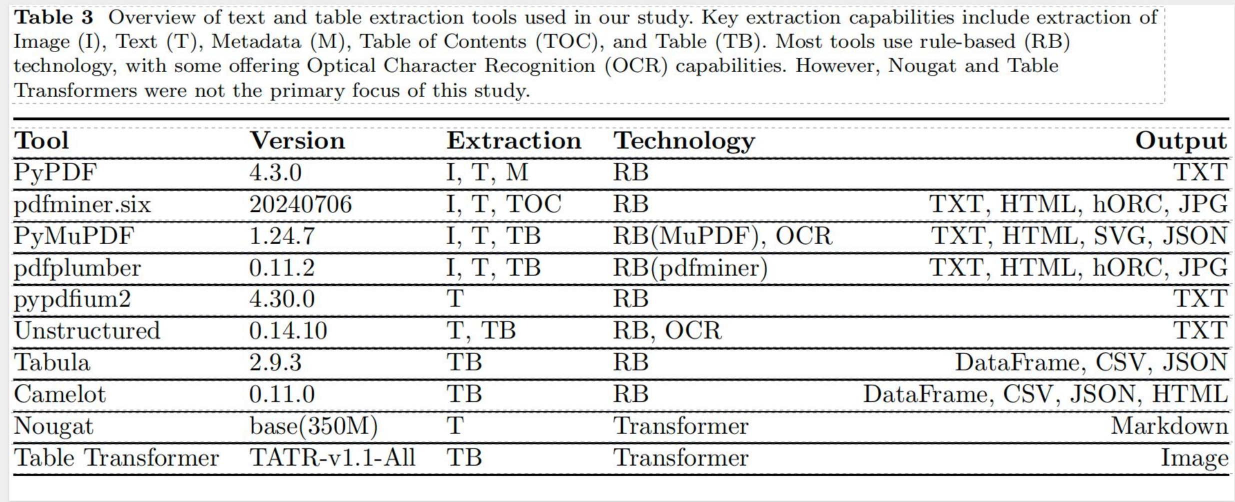

Table 3 Overview of text and table extraction tools used in our study. Key extraction capabilities include extraction of Image (I), Text (T), Metadata (M), Table of Contents (TOC), and Table (TB). Most tools use rule-based (RB) technology, with some offering Optical Character Recognition (OCR) capabilities. However, Nougat and Table Transformers were not the primary focus of this study.

Tool	Version	Extraction	Technology	Output
PyPDF	4.3.0	I, T, M	RB	TXT
pdfminer.six	20240706	I, T, TOC	RB	TXT, HTML, hORC, JPG
PyMuPDF	1.24.7	I, T, TB	RB(MuPDF), OCR	TXT, HTML, SVG, JSON
pdfplumber	0.11.2	I, T, TB	RB(pdfminer)	TXT, HTML, hORC, JPG
pypdfium2	4.30.0	T	RB	TXT
Unstructured	0.14.10	T, TB	RB, OCR	TXT
Tabula	2.9.3	TB	RB	DataFrame, CSV, JSON
Camelot	0.11.0	TB	RB	DataFrame, CSV, JSON, HTML
Nougat	base(350M)	T	Transformer	Markdown
Table Transformer	TATR-v1.1-All	TB	Transformer	Image