

# Text File Processing Workflow

## Evaluation Results

### Information Extraction

 [Image]

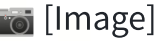
 [Image]

File Format	Tool	Dataset	Score	Extraction Time
PDF	MinerU	Self-constructed multi-type PDF files (total of 200 PDFs)	8.30 / 10	CPU / GPU: 1.47s/page
		Converted scanned PDFs to images	7.12 / 10	1.43s/page
	GPT-4o		9.25 / 10	—
Word	WordSplit	A self-created .docx file (~13,000 characters) with text, images, and embedded equations	Very accurate	0.12s / 13,000 chars
Excel	ExcelToJson	Self-created Excel files with nested tables, simple tables, and headerless tables	Very accurate; complex headers are harder to parse	0.019s / 472 cells

### AI Scoring Criteria (Information Extraction)

- 10: Perfect match, no meaningful error.
- 7–9: Minor formatting or OCR issues that do not affect comprehension.

- **4–6:** Moderate errors that impact accuracy or clarity.
- **0–3:** Major errors or large portions missing/incorrect.



## PDF Recognition Example

### Multimodal Document Understanding & Retrieval – Experimental Results

#### Image Understanding

Tool	Dataset (from PDF extractions)	Score	Extraction Time	
Qwen2.5-Omni-7B	852 self-extracted images from PDF dataset	8.04 / 10	2.31s/image	

#### AI Scoring Criteria (Image Understanding)

- **10:** Excellent description, detailed and fully accurate
- **7–9:** Good but missing minor points or has small issues
- **4–6:** Partial or vague understanding, moderate flaws
- **0–3:** Misleading, incorrect, or very low-quality result

## Text Embedding

### Embedding Performance

Model Name	Parameters	Dataset	MRR	Hit@5	GPU	CPU
inf-retriever-v1	7B	BEIR	0.7738	0.9167	A6000	AMD EPYC 965
inf-retriever-v1-1.5b	1.5B		0.7417	0.88	A6000	AMD EPYC 965
gte-Qwen2-1.5B-instruct	1.5B		0.6457	0.8	A6000	AMD EPYC 965
gte-Qwen2-7B-instruct	7B		0.7408	0.8566	A6000	AMD EPYC 965
Linq-Embed-Mistral	7B		0.7098	0.8333	A6000	AMD EPYC 965
jina-embeddings-v3	572M		0.5957	0.7266	A6000	AMD EPYC 965
gte-multilingual-base	305M		0.6134	0.76	A6000	AMD EPYC 965

## Embedding Efficiency