OCR Benchmark Test Document

1. English Text Section

Artificial intelligence (AI) has revolutionized document processing. Modern OCR systems leverage vision-language models (VLMs) to understand complex layouts including tables, mathematical formulas, and multilingual text. This document tests extraction accuracy across these dimensions.

2. Data Table

Model	Params (B)	Accuracy %	Speed (p/s)	Cost
MinerU 2.5	1.2	90.67	2.12	Free
DeepSeek-OCR-2	3.0	91.09	~2.3	$0.01/1K
GLM-OCR	0.9	94.62	1.86	$0.03/1M
Tesseract 5.x	N/A	~78.5	0.8	Free
GPT-4o Vision	~200	93.1	~1.0	$5/1M

3. Mathematical Formulas

- Euler's identity: e^(i*pi) + 1 = 0
- Quadratic formula: x = (-b +/- sqrt(b^2 - 4ac)) / (2a)
- Gaussian integral: integral(-inf, inf) e^(-x^2) dx = sqrt(pi)
- Bayes' theorem: P(A|B) = P(B|A) * P(A) / P(B)
- Cross-entropy loss: L = -sum(y_i * log(p_i))

4. Structured List

1. Layout detection and reading order preservation
2. Table structure recognition (rows, columns, merged cells)
3. Formula extraction to LaTeX notation
4. Multi-language text recognition (CJK, Arabic, Cyrillic)
5. Handwriting and low-quality scan handling

OCR Benchmark Test Document

5. Chinese Text Section

序基战

序基战

6. Financial Data Table

Quarter	Revenue	COGS	Gross Margin	OpEx	Net Income
Q1 2024	$12.5M	$4.2M	66.4%	$5.1M	$3.2M
Q2 2024	$14.8M	$4.9M	66.9%	$5.4M	$4.5M
Q3 2024	$16.2M	$5.3M	67.3%	$5.8M	$5.1M
Q4 2024	$18.9M	$6.1M	67.7%	$6.2M	$6.6M
FY 2024	$62.4M	$20.5M	67.1%	$22.5M	$19.4M

7. Code Snippet

```
def extract_text(pdf_path: str) -> dict:
    """Extract structured content from PDF."""
    from mineru import MinerUParser

    parser = MinerUParser(backend="vlm-http-client")
    result = parser.parse(pdf_path)

    return {
        "text": result.get_text(),
        "tables": result.get_tables(format="html"),
        "formulas": result.get_formulas(format="latex"),
        "images": result.get_images(),
    }
```

8. Special Characters & Edge Cases

Currency: $100.50 | EUR 85.30 | GBP 72.15

Email: test.user@example.com | Phone: +852-1234-5678

URL: https://github.com/opendatalab/MinerU

Fractions: 1/2, 3/4, 7/8 | Percentages: 99.97%

Date formats: 2024-12-25 | 25/12/2024 | Dec 25, 2024

Units: 100kg, 3.14m/s, 25C, 1024MB, 10Gbps