OCR Comprehensive Test Document

English · Chinese · Tables · Formulas · Code · Special Characters

1. English Text

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 is designed to test extraction accuracy across multiple dimensions simultaneously.

Key capabilities being tested:

1. Layout detection and reading order preservation
2. Table structure recognition (rows, columns, headers)
3. Mathematical formula extraction
4. Multi-language text recognition (English + Traditional Chinese)
5. Code block preservation with syntax
6. Special characters, currencies, and URLs

2. OCR Model Comparison

Model	Parameters	OmniDocBench Score	Speed (pages/s)	Cost
GLM-OCR	0.9B	94.62%	1.86	$0.03/1M tokens
DeepSeek-OCR-2	3.0B	91.09%	~2.3	$0.01/1K tokens
MinerU 2.5	1.2B	90.67%	2.12	Free (open source)
GPT-4o Vision	~200B	93.1%	~1.0	$5.00/1M tokens
Tesseract 5.x	N/A	~78.5%	0.8	Free (open source)

3. Financial Data

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**

4. Mathematical Formulas

- Euler's identity: $e^{(i\pi)} + 1 = 0$
- Quadratic formula: $x = (-b \pm \sqrt{(b^2 - 4ac)}) / 2a$
- Gaussian integral: $\int_{-\infty}^{\infty} e^{(-x^2)}\, dx = \sqrt{\pi}$
- Bayes' theorem: $P(A|B) = P(B|A) \cdot P(A) / P(B)$
- Cross-entropy loss: $L = -\Sigma_i\, y_i \cdot \log(p_i)$
- Softmax function: $\sigma(z_i) = e^{(z_i)} / \Sigma_j\, e^{(z_j)}$
- Matrix multiplication: $C = A \times B$, where $C_{ij} = \Sigma_k\, A_{ik} \cdot B_{kj}$

5. 中文測試 — 唐詩三首

靜夜思

【李白】

床前明月光，疑是地上霜。
舉頭望明月，低頭思故鄉。

登鸛雀樓

【王之渙】

白日依山盡，黃河入海流。
欲窮千里目，更上一層樓。

楓橋夜泊

【張繼】

月落烏啼霜滿天，江楓漁火對愁眠。
姑蘇城外寒山寺，夜半鐘聲到客船。

中文段落

人工智能技術正在深刻改變文檔處理領域。光學字符識別（OCR）系統已從傳統的基於規則的方法演進到基於深度學習的視覺語言模型。這些模型能夠理解複雜的文檔佈局，包括表格、公式和多語言文本。主要挑戰包括：手寫體識別、低質量掃描文檔處理、以及保持原始文檔的閱讀順序。

6. Code Snippet

```python
def extract_text(pdf_path: str) -> dict:
    """Extract structured content from PDF using GLM-OCR."""
    from gradio_client import Client, handle_file

    client = Client("prithivMLmods/GLM-OCR-Demo", token=HF_TOKEN)
    raw_output, rendered_md = client.predict(
        image=handle_file(pdf_path),
        task="Text",
        api_name="/process_image",
    )
    return {"raw": raw_output, "markdown": rendered_md}
```

7. Special Characters & Edge Cases

> Currency: $100.50 | €85.30 | £72.15 | ¥12,800 | HK$780.00
> Email: test.user@example.com | Phone: +852-1234-5678
> URL: https://github.com/opendatalab/MinerU
> Fractions: ½, ¾, ⅛ | Percentages: 99.97%
> Date formats: 2024-12-25 | 25/12/2024 | Dec 25, 2024

> Units: 100kg, 3.14m/s^2, 25°C, 1024MB, 10Gbps
>
> Symbols: © 2024 OpenClaw™ | α β γ δ ε | → ← ↑ ↓ | ★ ✓ ✗

8. 中英對照表

中文名稱	English Name	參數量	準確率
通義千問	Qwen-VL	7B	89.3%
深度求索	DeepSeek-OCR	3B	91.1%
智譜清言	GLM-OCR	0.9B	94.6%
書生浦語	InternVL	26B	92.8%