Text Extraction Research Report: OCR vs. API Extraction Models

📌 OCR & Document Extraction Test Summary (Generalized)

🖼️ Image OCR

- **Method:** EasyOCR
- **Average Time:** ~1.5 to 2 seconds
- **Performance:** Consistently accurate on small and clear images.
- **Preview Example:**

  photosynthesis process through which green plants create energy using carbon dioxide and water

---

📄 PDF Extraction

1. **Small PDF (1 page)**
   - **Method:** PDF Text Extraction
   - **Time Taken:** ~0.06 to 0.1 seconds
   - **Performance:** Extremely fast and accurate.
2. **Large PDF (~300+ pages)**
   - **Method:** PDF Text Extraction
   - **Time Taken:** ~4.4 to 5 seconds

  - **RAM Usage:** 30-57 MB
  - **Performance:** Stable and efficient.
  - **Note:** Gemini **could not** process PDFs of this size.

📝 DOCX Extraction

- **Method:** DOCX Extraction
- **Time Taken:** ~0.02 seconds
- **Performance:** Instantaneous and stable

🧩 ML Model System Summary

- **Capabilities:**
  - OCR, PDF extraction, **blocked PDF extraction**, DOC/DOCX support.
  - **Pagination features:** Measures and indexes each page.
  - **Font measurement:** Can identify and measure fonts, sizes, styles within pages.
  - Handles formats Gemini cannot: DOC, PPTX, XLSX.
- **Performance:**
  - **Much faster** than Gemini.
  - Uses **very low resources**.
- **Stability:**

    - Highly stable across various document types.
- **Requirement:**
    - Needs a **dedicated server** to function optimally.

⚠️ General Limitations

- **Gemini:**
    - No support for DOC, PPTX, XLSX.
    - Pagination and font detection are possible, but **unstable**.
    - Struggles beyond 15 pages or with complex formatting.
- **ML Model System:**
    - Requires server but compensates with **speed, stability, and wider format support**.

Small image ocr-

```
📊 --- Extraction Summary ---
📄 File: /content/tnt5mar87f (3).jpg
⚙️  Method Used: Image OCR (EasyOCR)
⏱️ Time Taken: 1.65 seconds
💾 RAM Used: 0.0 MB
📝 Saved Extracted Text To: /content/tnt5mar87f (3).jpg_extracted.txt

🔍 Preview:
photosynthesis process through which green plants create energy using carbon
dioxide and water
```

Small pdf

```
📊 --- Extraction Summary ---
📄 File: /content/hello (1).pdf
⚙️  Method Used: PDF Text Extraction
⏱️ Time Taken: 0.06 seconds
💾 RAM Used: 2.99 MB
📝 Saved Extracted Text To: /content/hello (1).pdf_extracted.txt

🔍 Preview:
=== Page 1 ===
hello dwdwq asdasdd as d asd asd asd das asd asd dsada s
```

Small docs

```
📊 --- Extraction Summary ---
📄 File: /content/hello (3).docx
⚙️  Method Used: DOCX Extraction
⏱️ Time Taken: 0.02 seconds
💾 RAM Used: 0.0 MB
📝 Saved Extracted Text To: /content/hello (3).docx_extracted.txt
```

```
📊 --- Extraction Summary ---
📄 File: /content/Adabiyot 9-sinf.pdf
⚙️  Method Used: PDF Text Extraction
⏱️ Time Taken: 4.41 seconds
💾 RAM Used: 30-57 MB
📝 Saved Extracted Text To: /content/Adabiyot 9-sinf.pdf_extracted.txt
```

Gemini Model Analysis

PDF OCR / Text Extraction (~10-15 pages max)

- **Time**: Scales roughly to 10-15 seconds.
- **Performance**: Good for text extraction and OCR in smaller PDFs.
- **Limitations**: Cannot reliably handle PDFs beyond ~15 pages.
- **Special Features**: Can sometimes detect pagination, fonts, and structure, but this is unstable and inconsistent.

Mini Image OCR

- **Model**: Gemini2-2.40
- **Time**: ~2 - 2.5 seconds
- **Performance**: Good accuracy with occasional instability.

Small PDF (1 page)

- **Time**: ~2.45 seconds

Gemini testing model
137 seconds, 11 pages

Mini image
Gemini2-2.40, good, unstable

Small pdfs:
2.45 seconds, good accuracy

No doc, no pptx, no xlsx, and sometimes not stable but when document has unexpected formats like bold spain or so, it can do it independently

Mistral Small 3 Analysis

Evaluation Criteria

Criteria	Evaluation
🧠 **Accuracy**	✅ **High** — Captures hourglass structure, intro/body/conclusion components, and cultural variation. Good term definitions and relevant comprehension questions.
⏱ **Time Taken**	⏳ **Moderate** — Slight delay, possibly due to rate limit, but fast once available.
💾 **Memory Usage**	⚡ **Lightweight on user side** — Processing occurs on OpenRouter's servers, requiring only prompt submission.
📐 **Output Structure**	✅ **Very Clear** — Cleanly separated into sections (summary, questions, terms) with effective use of bolding and bullet points.
🌍 **Understanding of Educational Task**	✅ Yes — Content structured like a study guide: informative, extractive, and well-labeled.

Conclusion

Gemini performs well for smaller PDFs and images with quick processing times but struggles with larger documents and inconsistent structural detection. Mistral Small 3 excels in accuracy and output structure, making it suitable for educational tasks, though it may experience slight delays.