PDF Processing Tools Evaluation

This document tests three PDF processing tools:

1. PyMuPDF - Fast text extraction
2. Marker - PDF to Markdown conversion
3. Nougat - Neural OCR for academic papers

Key Features

PyMuPDF:
- Lightweight and fast
- Works on CPU only
- Suitable for simple text extraction

Marker:
- Deep learning based
- Produces markdown output
- Good for RAG applications

Nougat:
- Academic paper specialist
- Handles mathematical equations
- Requires GPU for practical use

Conclusion

Each tool has its strengths. PyMuPDF is best for speed, Marker excels at structure preservation, and Nougat specializes in academic content with complex formatting.