# STEP 1: PARSE RESEARCH PAPERS (VIA PyMuPDF)

Before doing anything smart, we need to **extract the raw text** from the PDFs of research papers.

**Goal**: Convert a research paper into raw text (title, abstract, paragraphs, sections).

I am working in a python venv (virtual environment)

A virtual environment is created on top of an existing Python installation, known as the virtual environment’s “base” Python, and may optionally be isolated from the packages in the base environment, so only those explicitly installed in the virtual environment are available.

First I install PyMuPDF from pip.

Then I go through the basics like

1. print(fitz.\_\_doc\_\_) //PyMuPDF 1.26.0: Python bindings for the MuPDF 1.26.1 library (rebased implementation).

Python 3.13 running on win32 (64-bit).

1. pdf = fitz.open(‘a.pdf’) //opens the pdf which is present in the directory im running the terminal from
2. print(pdf.page\_count) //prints no of pages of pdf
3. print(pdf.metadata[‘author’]) //prints meta data of pdf
4. print(pdf.get\_toc()[:10]) //prints first 10 table of contents of pdf
5. page = pdf.load\_page(9) //loads the given page, remember indexing starts from 0 not 1
6. text = pdf.get\_text(‘text’) //extracts the text into ‘text’, can then print(text)
7. print(text.replace(‘\t’, ‘\n’) //replaces from the text