


# AI/Python Intern Assignment: PDF Content Analysis and Question Generation

## Introduction

Welcome! This assignment is designed to assess your skills in Python, AI/ML, and application development. The goal is to build a system that can process a PDF document, extract its contents (text and images), and then use AI to generate questions based on the visual information.

For this assignment, you will be working with the provided sample file:

 IMO class 1 Maths Olympiad Sample Paper 1 for the year 2024-25.pdf

## Project Overview

You are tasked with creating a Python-based tool that can analyze a PDF file, specifically one containing educational content like the sample provided. The project is divided into three main parts.

### Part 1: PDF Content Extraction

Your first task is to extract all meaningful content from the PDF.

#### Requirements:

- Text Extraction:** Write a Python script that extracts all textual content from the IMO Grade 1 - 1-2.pdf file.
- Image Extraction:** Your script should also identify and extract all images from the PDF. Save these images as separate files (e.g., page1\_image1.png, page2\_image1.png).
- Structured Output:** Create a JSON file that organizes the extracted content. The JSON should have a structure that lists the content page by page, including the text and the file paths to the images found on each page.

#### Example JSON Structure:

```
[
  {
    "question": "What is the next figure?",
    "images": "path/to/question_image_1.png",
    "option_images": ["path/to/option_image_1.png", "path/to/option_image_2.png"]
  },
  {
```

```
"question": "What is the next figure?",  
"images": ""path/to/question_image_1.png",  
"option_images": ["path/to/option_image_1.png", "path/to/option_image_2.png"]  
},  
]
```

**Libraries you might find useful:** PyMuPDF (or fitz), pdfplumber, Pillow.