The aim of this study is to develop an automated and scalable PDF text extraction system using Python, incorporating Optical Character Recognition (OCR) and advanced parsing techniques to accurately process diverse document structures while ensuring efficiency, reliability, and usability