

The Challenge : The Document Data Bottleneck

Businesses worldwide struggle with unstructured PDF data. Extracting information—from complex financial statements to specific clauses in legal documents—is slow, manual, and error-prone. Furthermore, table recognition creates further bottlenecks. These bottlenecks hinder automation and prevent modern AI models from efficiently utilizing vast amounts of valuable document data.

Our Solution: Open Data Loader PDF

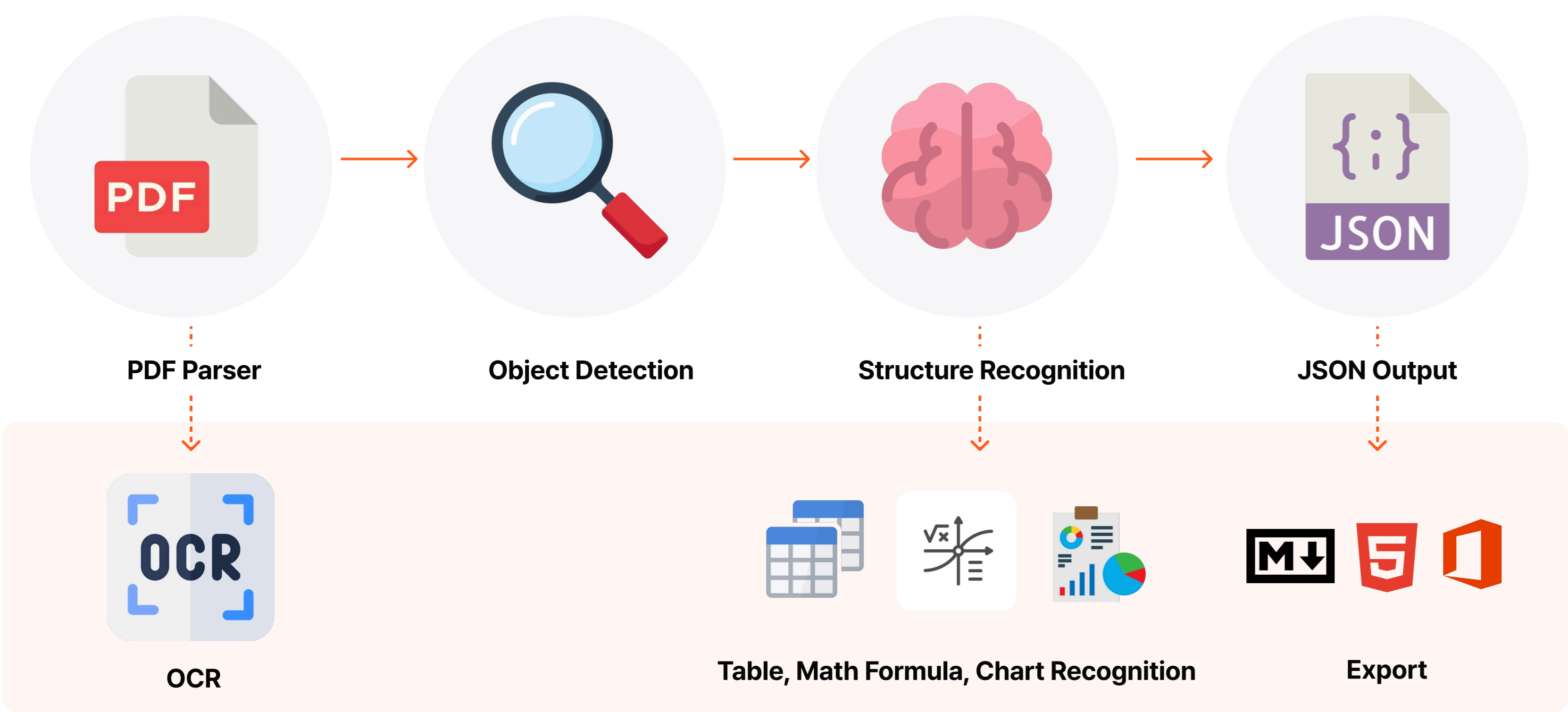
Open Data Loader PDF is a high-performance open-source **SDK engine** that addresses this challenge head-on. Developed in partnership between HANCOM and Dual Lab, our engine transforms unstructured PDFs into clean, structured data, making it the perfect foundation for **AI-driven workflows**. We are building the **world's fastest** and most accurate PDF data extraction engine.

Key Features : The fastest Data Loader PDF

- ⚡ **Fast** – Efficient batch processing for thousands of documents.
- ✅ **Accurate** – Extracts text, tables, images, and layout with high precision.
- 🔒 **Secure** – Runs fully offline, making it ideal for sensitive or regulated environments.
- 🧠 **AI-Ready** – Provides structured outputs in Markdown or JSON, optimized for LLMs.
- 🔧 **Customizable** – Rule sets can be adapted for domain-specific structures.
- 🔗 **AI Add-on Compatible** – Seamlessly upgrade to AI-powered modules when needed.

Technical Strategy

Our engine's architecture is a streamlined workflow designed for clarity and efficiency. The diagram below illustrates how we process documents from a raw PDF to a structured JSON output, with a focus on our core capabilities.



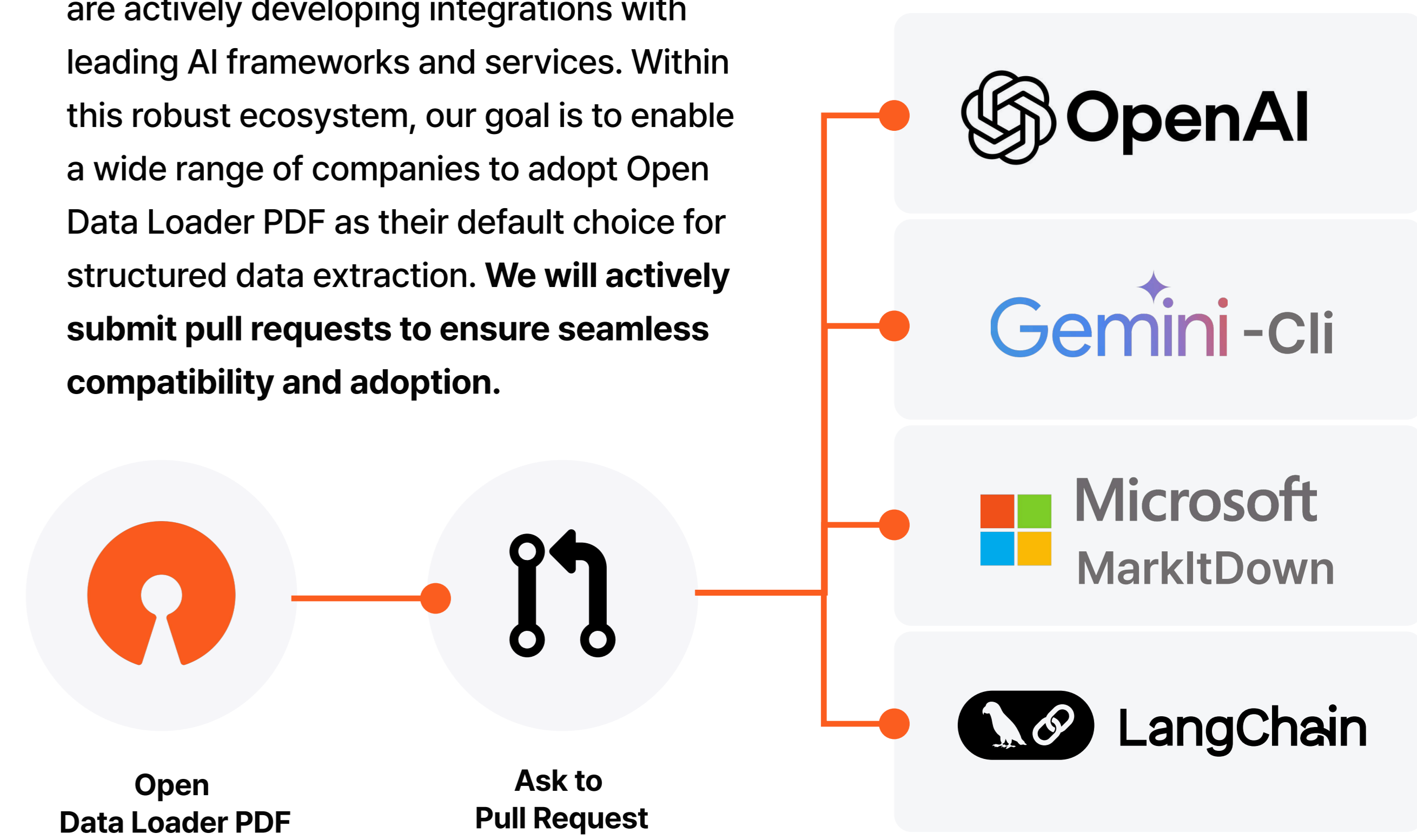
Use Cases & Success Scenerio

Our SDK is designed to solve real-world problems and foster developer innovation across a wide range of industries. It can be used in the following industries:

- Financial Services**
Automate the extraction of structured data from invoices, financial statements, and annual reports for real-time analysis and reporting.
- Legal & Compliance**
Rapidly parse contracts and legal documents to identify key clauses, dates, and entities, accelerating due diligence and review processes.
- Research & Academia**
Extract tables and figures from scientific papers and research documents, enabling automated data collection and meta-analysis.
- Enterprise Document Automation**
Build custom document processing workflows to convert legacy PDFs into modern, searchable, and machine-readable formats.

AI integration Strategy

To accelerate the AI-based ecosystem, we are actively developing integrations with leading AI frameworks and services. Within this robust ecosystem, our goal is to enable a wide range of companies to adopt Open Data Loader PDF as their default choice for structured data extraction. **We will actively submit pull requests to ensure seamless compatibility and adoption.**



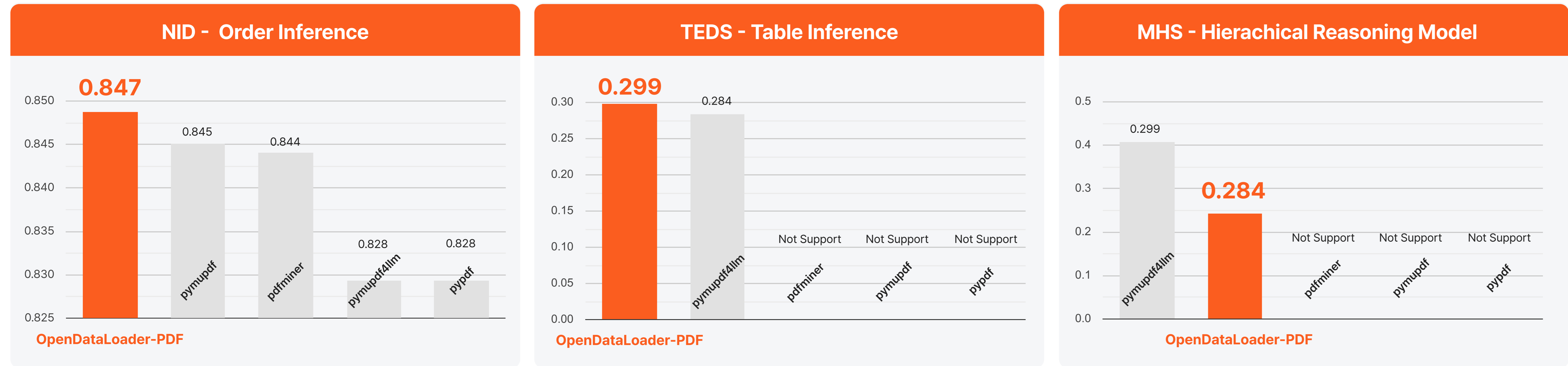
Benchmark Evaluation: Performance Validation of Open Data Loader PDF

The Open Data Loader PDF was evaluated using a benchmark method called Document Parsing Benchmark across three key metrics: NID - Order Inference, TEDS - Table Inference, and MHS - Hierarchical Reasoning Model. All evaluation metrics were benchmarked against Rule-Based engines.

The Document Parsing Benchmark is composed of a collection of 200 benchmark samples, meticulously curated from various public datasets (KLEISTER-NDA, PubLayNet, CORD, FUNSD, DocVQA) to comprehensively assess Document Understanding capabilities. This approach focuses on objectively validating the engine's performance in structure recognition and data extraction across a diverse range of document types and complexities.

We Love Open Source

Hancom and Dual Lab are working in close collaboration to successfully lead the Open Data Loader PDF project. By combining our deep expertise in document processing technology and AI, we are committed to enhancing the project's technical excellence. Through our continued cooperation, we will strive to create a variety of successful use cases and innovative applications for the Open Data Loader PDF, bringing new vitality to the open-source ecosystem.



HANCOM × dual lab

OpenDataLoader-PDF : GitHub Page

Please Visit And Join the Journey with Us!

Your interest and feedback are invaluable to us. Come explore our code, open issues and become part of our growing community

Visit us | github.com/opendataloader-project
Contact us | open.dataloader@hancom.com

© 2025 Hancom Inc. All rights reserved.

