Inclusive Development International (IDI) monitors environmental and social harms in global palm oil supply chains through PalmWatch, an open-access platform that maps brands to individual mills. However, the data PalmWatch needs is often published by brands as inconsistently formatted PDF disclosures, with borderless tables, tiny cells, and varying layouts, making large-scale supply chain monitoring difficult. To address this, the Data Science Clinic partnered with IDI to develop an automated pipeline that converts these disclosures into standardized datasets (Figure 1) containing mill names, parent companies, certification status, and geographic data.

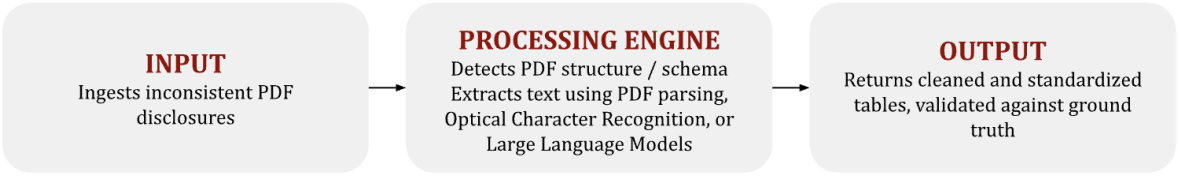

Figure 1: Pipeline for extracting mill-level supply chain information from inconsistent PDF disclosures

The team evaluated multiple approaches, including rule-based parsers, Optical Character Recognition (OCR) techniques, and Large Language Model (LLM) methods. Each approach was benchmarked on extraction accuracy, latency, cost, robustness across different brand disclosures, and ease of maintenance (Figure 2).

The recommended pipeline, a rule-based parser with an LLM fallback, achieved the highest average column accuracy of 77.3%, while processing each brand in average 3.5 minutes using moderate compute resources (~23k input / 0.15k output tokens). It also showed moderate robustness across disclosure formats and required less manual configuration than competing methods. This makes it the most scalable solution for helping IDI monitor supply chain-related environmental and social harms.

Method	Average Column Accuracy	Cost ($)	Latency / company (min)	Robustness	Maintainability
Rule-Based Parsing with LLM Fallback	*77.3%*	*Medium: 0.15k output & 23k input token usage*	*~ 3.5*	*Medium*	*High*
OCR-Based Parsing with Human Feedback	75%	Medium: GPU compute cost	CPU: ~50 GPU: ~5–10	Medium	Low
Direct LLM Approach	36%	High: 65k output & 65k input token usage	~ 60	Medium	Low

Figure 2: Comparison of PDF extraction approaches across various key dimensions