Evaluation of IDP through Hybrid Extraction: Merging LLMs, ML, and Layouts

# Abstract

Intelligent Document Processing (IDP) is crucial for converting unstructured documents into structured, actionable data. This paper presents an evaluation of a hybrid extraction framework that integrates large language models (LLMs) with deterministic methods, traditional machine learning and deep learning (ML/DL) techniques, and layout-based approaches. By combining semantic extraction from LLMs, reliable output from deterministic locators using OCR positional data, and lightweight ML/DL models for structured tasks (such as address segmentation), the proposed framework enhances robustness, reproducibility, and cost-efficiency in production environments. An automated evaluation pipeline cross-validates outputs from the different modules and continuously refines the system through a feedback loop. Experimental results on diverse document types demonstrate that the hybrid approach outperforms single-method solutions in accuracy and reliability. The paper discusses system design, deployment challenges, and future directions to further optimize IDP in real-world applications.

# 1. Introduction and Motivation

The increasing complexity and variability of real-world documents have underscored the importance of robust Intelligent Document Processing (IDP). Traditional extraction methods alone often fail to handle diverse formats or compensate for OCR errors. Our hybrid approach integrates LLMs for capturing semantic nuances, deterministic methods leveraging OCR positional data, and lightweight ML/DL models for tasks such as address parsing. This combination provides a more reliable, reproducible, and cost-effective solution for production environments.

# 2. Related Work

Recent work in document processing has demonstrated the benefits and limitations of LLM-based extraction, deterministic, layout-based approaches, and traditional ML/DL techniques. However, no single method fully addresses the challenges of robust production-level IDP. Our approach aims to bridge this gap by merging the strengths of these methods.

# 3. Proposed Hybrid Extraction Framework

Our proposed framework comprises three parallel pipelines:  
- \*\*LLM Module:\*\* Uses large language models to extract semantic and contextual information from text.  
- \*\*Deterministic/Layout Module:\*\* Employs OCR outputs and rule-based locators to capture document layout and structure.  
- \*\*ML/DL Module:\*\* Applies lightweight, non-generative transformer models and other traditional ML/DL methods for structured tasks such as address segmentation.  
  
The outputs from these modules are fused using a strategy (e.g., weighted averaging or rule-based reconciliation) to generate a robust final extraction. An automated evaluation pipeline then cross-validates these outputs, and any discrepancies trigger a feedback loop for continuous system refinement.

# 4. Automated Evaluation and Continuous Refinement

We implement an automated evaluation pipeline that monitors both offline metrics (precision, recall, F1 score) and production metrics (latency, consistency, error rates). This system automatically compares outputs from the different extraction pipelines, identifying inconsistencies and triggering retraining or fine-tuning when necessary. The feedback loop ensures continuous improvement and adaptation to new document formats and business requirements.

# 5. Experimental Setup and Results

Our experimental evaluation uses a diverse set of documents including invoices, contracts, academic papers, and addresses. We detail the technical setup (models, libraries, and infrastructure) and compare our hybrid approach against baseline single-method solutions. Results show improvements in accuracy (e.g., higher F1 scores), robustness (consistent outputs across runs), and resource efficiency (lower computational overhead). Tables and figures illustrate these gains.

# 6. Discussion

The hybrid extraction framework effectively combines the strengths of LLMs, deterministic locators, and lightweight ML/DL methods, resulting in a system that is more robust and reproducible in production. We discuss practical implications, scalability, and cost-effectiveness, as well as limitations such as increased system complexity and integration challenges. Future research directions include exploring additional fusion strategies and further automating the feedback loop.

# 7. Conclusion and Future Work

Our work demonstrates that integrating LLMs with deterministic and ML/DL-based extraction techniques significantly improves IDP performance in production environments. The hybrid framework not only enhances accuracy and robustness but also provides a cost-efficient and scalable solution. Future work will focus on optimizing fusion strategies, enhancing the automated evaluation pipeline, and extending the approach to additional document domains and languages.

# 8. Practical Deployment Considerations

We discuss strategies for integrating the hybrid framework into existing production pipelines, including continuous monitoring, maintenance, and scaling methods. Recommendations for industry practitioners on handling high-volume document processing and managing computational resources are also provided.

# References

Include relevant citations from recent work on hybrid extraction, LLM-based extraction, deterministic methods, and lightweight ML/DL techniques.