Document Extraction Benchmark Report

Performance Analysis Q4 2025

Executive Summary

This report presents a comprehensive analysis of document extraction libraries tested across multiple document formats and complexity levels. Our evaluation focused on extraction accuracy, processing speed, and resource utilization.

Test Results

Performance Comparison

Library	Avg Time (s)	Success Rate	Text Quality
Unstructured	1.23	98%	Good
Azure DI	2.45	99%	Excellent
PyPDF2	0.87	85%	Fair
PyMuPDF	0.95	92%	Good

Recommendations

1. Use Unstructured for cost-sensitive applications with simple documents

2. Deploy Azure DI for production systems requiring high accuracy

3. Implement hybrid approach for optimal cost-performance balance

4. Monitor extraction quality metrics continuously

Conclusion

Our analysis demonstrates that no single library is optimal for all use cases. The choice depends on specific requirements including budget constraints, accuracy requirements, and document complexity. A hybrid approach combining multiple libraries based on document characteristics yields the best results.