Hybrid OCR-LLM Framework for Enterprise-Scale Document Information Extraction Under Copy-heavy Task

Zilong Wang
Ningbo Institute of Digital Twin
Eastern Institute of Technology
Ningbo, Zhejiang 315200, P.R. China
zlw@idt.eitech.edu.cn

Xiaoyu Shen*
Ningbo Institute of Digital Twin
Eastern Institute of Technology
Ningbo, Zhejiang 315200, P.R. China
xyshen@eitech.edu.cn

Abstract

Information extraction from copy-heavy documents, characterized by massive volumes of structurally similar content, represents a critical yet understudied challenge in enterprise document processing. We present a systematic framework that strategically combines OCR engines with Large Language Models (LLMs) to optimize the accuracy-efficiency trade-off inherent in repetitive document extraction tasks. Unlike existing approaches that pursue universal solutions, our method exploits document-specific characteristics through intelligent strategy selection. We implement and evaluate 25 configurations across three extraction paradigms (direct, replacement, and table-based) on identity documents spanning four formats (PNG, DOCX, XLSX, PDF). Through table-based extraction methods, our adaptive framework delivers outstanding results: F1=1.0 accuracy with 0.97s latency for structured documents, and F1=0.997 accuracy with 0.6 s for challenging image inputs when integrated with PaddleOCR, all while maintaining sub-second processing speeds. The 54× performance improvement compared with multimodal methods over naive approaches, coupled with format-aware routing, enables processing of heterogeneous document streams at production scale. Beyond the specific application to identity extraction, this work establishes a general principle: the repetitive nature of copy-heavy tasks can be transformed from a computational burden into an optimization opportunity through structure-aware method selection.

1 Introduction

Copy-heavy tasks, characterized by large-scale processing of highly repetitive and template-based documents, pose persistent challenges in enterprise environments. From insurance claims and government forms to financial reports and identity documents, as shown in Figure 1, enterprises routinely extract structured information from millions of similar documents daily (Gagie et al., 2017; Navarro, 2019; Cobas and Navarro, 2019). While the structural redundancy offers potential for optimization, it also exacerbates issues of computational inefficiency, error propagation, and system brittleness.

Traditional extraction systems rely heavily on rule-based or template-specific configurations (Majumder et al., 2020; Xu et al., 2020). These approaches often perform well under controlled conditions but degrade rapidly with minor format changes, limiting their scalability across diverse enterprise settings (Gunel et al., 2022; Zhang et al., 2024). Moreover, quality assurance remains difficult without labeled benchmarks (Seitl et al., 2024), and integration with downstream systems is often ad hoc and fragile (Tang et al., 2021).

Large Language Models (LLMs) offer new possibilities for zero-shot and instruction-based document understanding. However, their practical use for structured extraction remains limited by high latency, hallucination risks (Li et al., 2024), and inefficiencies in handling repetitive or low-variance content. In copy-heavy tasks, latency is especially pronounced because generative models must decode output token by token; what an OCR/post-processing stack can copy in (near) constant time becomes hundreds or thousands of sequential generations. This generation-first workflow wastes time and budget on text that could be copied verbatim and increases exposure to stochastic errors. In production pipelines that demand sub-second response times and high precision—such as identity information extraction—these limitations can lead to unacceptable failure rates (Cooney et al., 2023).

Identity document processing exemplifies the core challenges of copy-heavy extraction. Fields such as names, dates, and ID numbers appear in consistent formats across documents, yet achieving 100% accurate, high-throughput extraction remains elusive. Minor OCR errors or model inconsisten-

*Corresponding Author

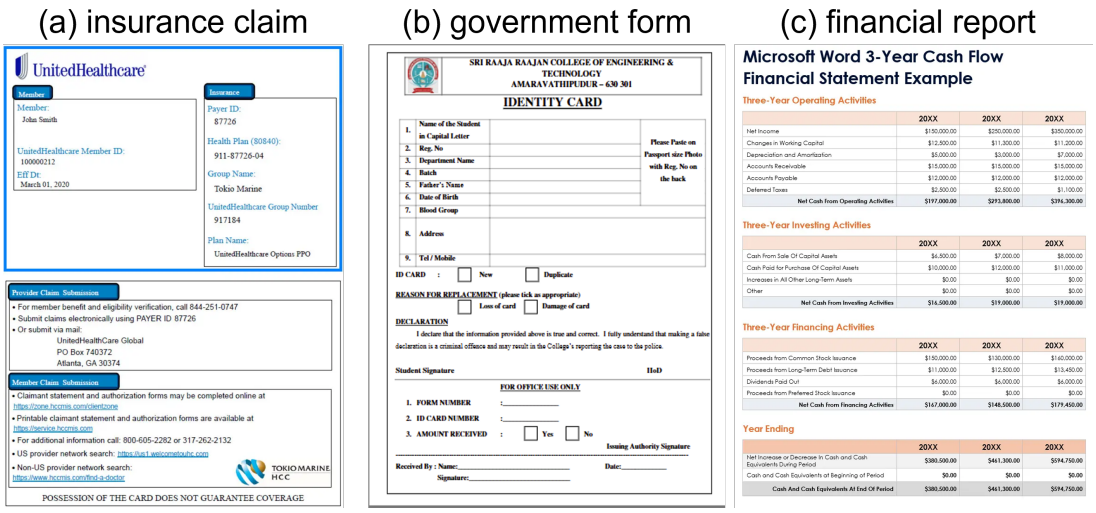

Figure 1: Document examples for copy-heavy tasks: (a) insurance claim; (b) government form; (c) financial report.

cies can introduce systematic faults across entire datasets, highlighting the need for robust, adaptive extraction frameworks.

In this paper, we propose a fast and adaptive multi-method extraction framework that integrates traditional OCR engines with LLM-based strategies to process copy-heavy documents efficiently. Our contributions are as follows.

- **A hybrid extraction framework** combining multiple OCR engines with four LLM-based paradigms—Direct, Replace, Table, and Multimodal—tailored to different document structures and modalities.
- **A diverse empirical evaluation** of 25 method combinations over documents spanning multiple formats (e.g., PNG, DOCX, XLSX, PDF), uncovering strategy-specific trade-offs in speed and accuracy.
- **A document-aware method selection strategy** that achieves perfect F1 scores (1.000) with sub-second latency (0.97s average) by matching extraction methods to document characteristics.
- **Practical deployment insights** including guidance on OCR engine selection, integration challenges, and system-level design for scalable enterprise deployment.

Our work bridges the gap between LLM capabilities and production demands, providing a practical pathway for deploying high-performance extraction systems across repetitive document tasks in real-world enterprise settings.

2 Related Work

Traditional Paradigms in Document Information Extraction. Early systems for information extraction (IE) from structured or semi-structured documents were predominantly rule-based or template-driven (Appelt, 1999; Chiticariu et al., 2013). These methods offered high precision and interpretability without the need for annotated training data. However, they were notoriously brittle—any minor variation in layout, field order, or formatting could break the rules, leading to maintenance-heavy pipelines that scale poorly (Li et al., 2020). To reduce rule authoring overhead, early work explored template induction (Anick and Flynn, 1992; Freitag, 2000), and more recently, TWIX (Lin et al., 2025) exploits redundancy across similar documents to automate extraction. Nonetheless, the lack of robustness to visual drift and document variation remains a limitation (Majumder et al., 2020; Wang et al., 2020).

To address annotation costs, unsupervised techniques have been explored, such as pattern mining (Etzioni et al., 2005) and structural clustering (Allahyari et al., 2017), which infer field patterns across documents without labels. However, these approaches often yield noisy outputs and require significant post-processing (Riedel et al., 2010). Feature-based supervised models (Finkel et al., 2005; Ratinov and Roth, 2009) introduced stronger learning capacity, but still relied heavily on handcrafted features and domain-specific engineering, with limited generalizability to new document layouts or types (Yadav and Bethard, 2019).

Neural and Pretrained Architectures for Doc-

ument Understanding. The advent of deep learning introduced a major shift. Sequence models like BiLSTM-CRF (Lample et al., 2016; Ma and Hovy, 2016; Shen et al., 2017) enabled end-to-end learning for token classification tasks, significantly reducing manual feature design. These were soon outpaced by pre-trained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019; Su et al., 2022), which provided strong contextual embeddings for text-based extraction. However, most PLMs ignore document structure, which is critical in forms, invoices, and tables.

To incorporate visual and spatial cues, layout-aware PLMs such as LayoutLM (Xu et al., 2020), LayoutLMv2/v3 (Huang et al., 2022), and FormNet (Lee et al., 2022) combine text, layout, and visual embeddings. More sophisticated multimodal transformers like DocFormer (Appalaraju et al., 2021) further improve modality fusion and downstream accuracy. These models show strong performance on visually complex tasks but typically require finetuning and careful preprocessing (e.g., OCR + bounding boxes), and remain sensitive to layout noise and OCR errors.

LLMs and Prompt-based Extraction. Large Language Models (LLMs) have introduced powerful new paradigms for document IE, particularly in zero-shot and few-shot settings. Prompt-based models like GPT (Wang et al., 2023b; Wei et al., 2023) and InstructUIE (Wang et al., 2023c) frame extraction tasks as natural language instruction following. Methods such as self-prompting (Li et al., 2022) and chain-of-thought reasoning (Wei et al., 2022) improve consistency and reasoning over complex inputs. Architecturally, two key approaches emerge: (1) OCR-to-LLM pipelines like LMDX (Perot et al., 2023) and DocLLM (Wang et al., 2023a), which linearize OCR text into prompts for token-level generation, and (2) multimodal LLMs (Kim et al., 2022; Tang et al., 2023) that directly process document images, bypassing OCR and jointly modeling layout and content.

While these models are flexible and powerful, they pose several practical challenges: inference latency, high compute cost, difficulty in debugging hallucinations (Li et al., 2024; Zhang et al., 2023a), and limits in handling highly repetitive or deterministic document structures (Cooney et al., 2023). Furthermore, most current benchmarks (e.g., FUNSD (Jaume et al., 2019), DocVQA (Mathew et al., 2021)) emphasize semantic diversity and visual clutter, underrepresenting production settings dominated by redundant, high-volume forms.

Copy-heavy Extraction: A Distinct and Underexplored Setting. In many industrial scenarios, document extraction is not about semantic diversity but rather **structural redundancy**: processing millions of near-identical layouts (e.g., utility bills, ID cards, customs forms). We refer to this as **copy-heavy extraction**. In these settings, speed, fault tolerance, and robustness to minor OCR/model variance outweigh the need for general reasoning or layout flexibility. Even small errors can cascade into critical downstream failures in enterprise systems.

Existing LLM and multimodal models are often over-engineered for such tasks, leading to inefficiencies. Meanwhile, rule-based or template systems are fast but brittle. This tension between generality and efficiency is not well captured in most academic benchmarks, and the trade-offs in copy-heavy settings remain poorly studied (Seitl et al., 2024). These tasks demand tailored strategies that optimize for high-throughput, low-latency, and graceful degradation in the presence of noise.

Scalability Challenges and Emerging Optimizations. Recent work has started addressing these issues from a systems perspective. Techniques like batch prompting (Cheng et al., 2023), prefix sharing (BatchLLM (Zheng et al., 2024)), and intermediate caching (PromptCache (Gim et al., 2024), AttentionStore (Zhang et al., 2023b)) reduce redundant computation in LLM-based pipelines, particularly when input documents are similar. Lightweight architectures such as Donut (Kim et al., 2021) and UDOP (Tang et al., 2023) offer OCR-free alternatives that retain structured output formats. However, these models still suffer from trade-offs in latency, model size, and control granularity—key considerations for real-world deployments. Industrial settings also require interoperability with diverse file formats (PDF, DOCX, scanned images), graceful fallback mechanisms, and integration with legacy systems.

Toward Modular, Hybrid, and Document-aware Pipelines. Copy-heavy document extraction tasks—common in industrial settings like invoices, IDs, and utility forms—prioritize speed, robustness, and consistency over general reasoning. Existing solutions either rely on brittle rules or overgeneralized LLMs, leading to inefficiencies in scalability, latency, or maintainability.

We propose a modular and hybrid pipeline that

integrates OCR fusion, lightweight heuristics (e.g., field substitution, table-position extraction), and selective LLM invocation. A centralized, document-aware controller dynamically orchestrates these components based on document format, enabling fast, accurate, and scalable extraction. Unlike prior approaches that treat components in isolation, our system emphasizes integration and operational efficiency for real-world deployment.

3 Methodology

3.1 System Architecture

Our extraction framework comprises two primary components: text extraction and LLM-based information extraction. Figure 2 presents an overview of the complete pipeline. Initially, various tools are employed to extract textual content from heterogeneous file formats, including Markdown, Word, Excel, PDF, and images. The extracted text is then passed to a large language model (LLM) to perform task-specific information extraction.

To support different document structures and information needs, we implement multiple extraction strategies. Specifically, our methods fall into three categories, as shown in Figure 3: direct extraction, replacement-based extraction, and table-based extraction, which will be described in detail later. Additionally, we incorporate a multimodal model to directly extract information from images, serving as a baseline for comparison with text-based pipelines.

3.2 Text Extraction Tools

The quality of information extraction heavily depends on the accuracy and completeness of upstream text extraction. To support diverse input formats, we integrate a suite of specialized extraction tools, each selected based on its ability to handle specific file types and structural characteristics.

MarkItDown serves as a general-purpose parser for structured text formats, supporting Markdown, Word, Excel, and standard PDF documents. It preserves key structural features such as headings, tables, and semantic markers, enabling clean and semantically meaningful text extraction across multiple formats.

Docling is applied to both Word and PDF files, offering enhanced layout analysis and document hierarchy preservation. It is particularly effective in maintaining reading order and spatial layout, which are essential for downstream structure-aware tasks.

MinerU complements Docling by targeting complex PDF layouts, including multi-column formatting, dense tables, and embedded formulas. It provides fine-grained spatial structure recovery and is optimized for documents where layout fidelity is critical.

PaddleOCR and **EasyOCR** are employed for image-based documents. PaddleOCR is used as the primary engine due to its high accuracy in multilingual printed text, while EasyOCR acts as a fallback for cases involving handwritten content or degraded image quality.

The use of multiple tools for overlapping file types is intentional. Our goal is to systematically evaluate different extraction solutions on the same document type in order to identify the most effective extraction strategy. Evaluation focuses on two key dimensions: (1) the accuracy of extracted textual content, and (2) the fidelity of spatial and structural information preservation, which is crucial for position-aware downstream tasks such as table extraction or form understanding.

3.3 Target Information Extraction Methods

Once the raw text is extracted, we apply LLM-based methods to retrieve the target information. To accommodate different document layouts and content patterns, we design three complementary extraction strategies, each tailored to specific structural characteristics.

Direct extraction. This method applies LLMs or VLMs to perform end-to-end information extraction. In the text-based variant, raw text obtained from upstream extraction tools is passed to an LLM along with task-specific prompts. The vision-based variant bypasses OCR and directly uses document images as input to multimodal models, enabling better spatial understanding at the cost of higher inference overhead.

Replace extraction. To address the challenges of repetitive pattern documents, we adopt a two-step approach. First, structured elements, such as identifiers, are replaced with unique placeholders. The LLM is then prompted to retrieve associated fields based on these placeholders. This design improves consistency, reduces ambiguity, and allows efficient batch processing through prompt reuse.

Table extraction. For documents with tabular layouts, we combine LLM-based structure recognition with deterministic parsing. The LLM identifies table regions and target cell coordinates, while a rule-based parser extracts the content. This ap-

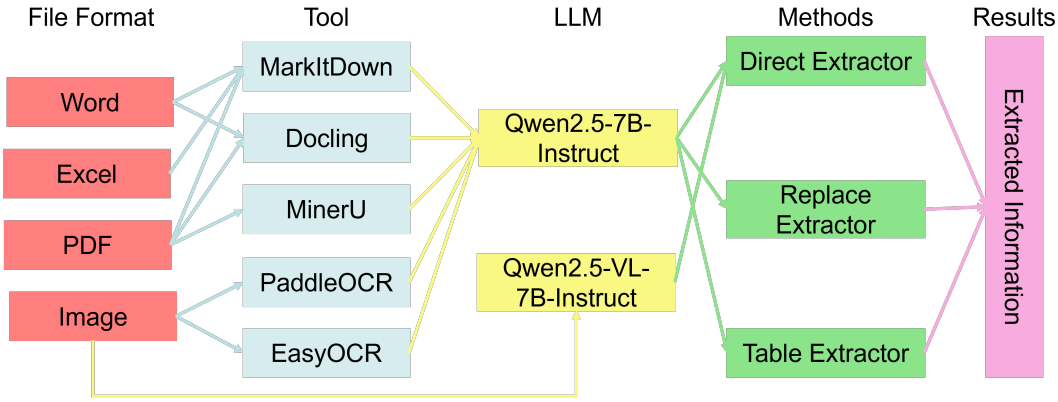

Figure 2: Architecture of the extraction framework showing the two main components: OCR processing, and LLM-based extraction.

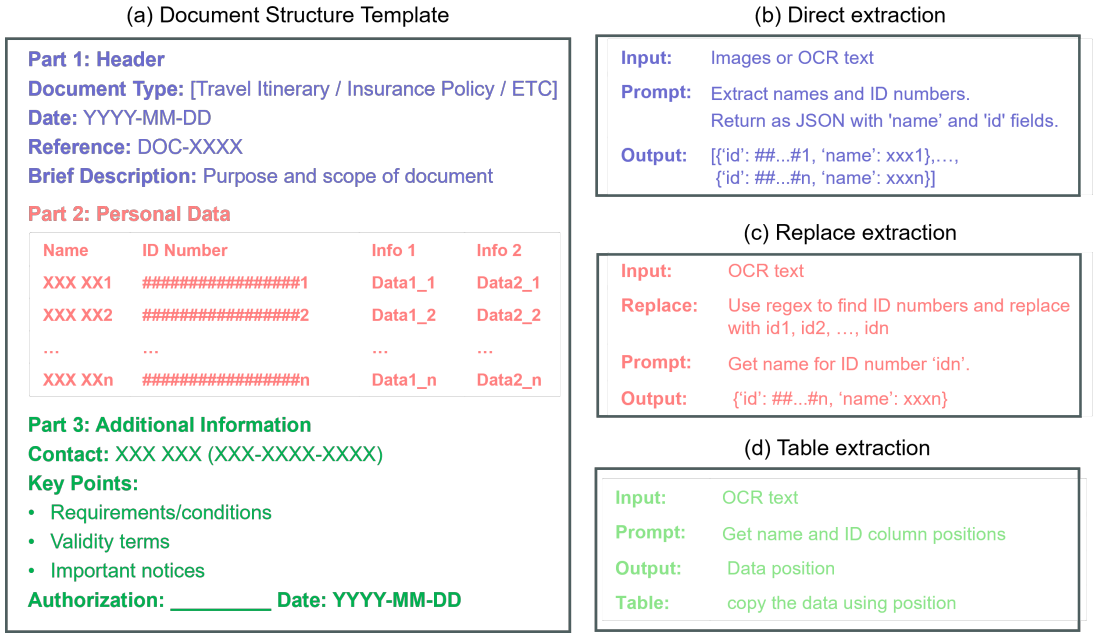

Figure 3: Extraction methods for copy-heavy tasks: (a) Document Structure Template; (b) Direct extraction; (c) Replace extraction; (d) Table extraction.

proach minimizes hallucination and reduces generation cost by limiting model output to positional metadata.

To determine the optimal strategy for different document types, we systematically compare the performance of these methods when paired with various text extraction tools. This enables us to identify the most effective combination in terms of both content accuracy and time consumption, ensuring adaptability and robustness in real-world document scenarios.

3.4 Model Selection

We adopt **Qwen2.5-7B** for text-based extraction and **Qwen2.5-VL-7B** for vision-language tasks, based on their strong balance between performance and efficiency. The 7B scale offers competitive extraction accuracy with significantly lower inference cost compared to larger models, enabling practical deployment on a single GPU.

Qwen2.5 models also provide robust multilingual support—particularly for Chinese, Japanese, and Korean—and share a unified architecture across text and vision variants, simplifying prompt design and enabling consistent evaluation across modalities. Their strong instruction-following ability ensures reliable structured outputs (e.g., JSON, placeholders), reducing post-processing overhead. These capabilities make them well-suited for our production-oriented information ex-

traction pipeline.

4 Experiments

4.1 Dataset

To evaluate our multi-method extraction framework, we construct a large-scale synthetic dataset of Chinese identity documents using GPT-4. The dataset simulates real-world diversity in format and content while preserving perfect ground truth.

Data Generation Pipeline. We generate 10–30 identity entries per document using the Faker library (zh_CN locale), producing realistic Chinese names and 18-digit ID numbers. These entries are embedded into semantically plausible documents (e.g., insurance forms, travel records, registration sheets) via GPT-4 prompts. Each document contains contextual information such as dates, headers, and auxiliary fields.

Document Formats. To evaluate cross-modality robustness, each generated Markdown document is converted into four formats:

- **PNG** (100): Rendered HTML to image (via `imgkit`), simulating scanned documents
- **DOCX** (100): Word documents with tables and formatted sections
- **PDF** (100): Generated using `WeasyPrint`, preserving layout and structure
- **XLSX** (100): Spreadsheets with tabular identity data

The final dataset comprises 400 documents with over 10,000 name-ID pairs.

4.2 Evaluation Metrics

We evaluate extraction performance at the (name, ID number) pair level.

Accuracy Metrics.

- **Precision:** Ratio of correctly extracted pairs to all extracted pairs $exactmatchrequired$
- **Recall:** Ratio of correctly extracted pairs to total ground truth pairs
- **F1 Score:** Harmonic mean of precision and recall

Efficiency Metrics.

- **Text Extraction Time:** Time for text extraction using each tool (MarkItDown, PaddleOCR, EasyOCR, etc.)
- **LLM Time:** Inference time for the extraction model (excluding text extraction)
- **Total Time:** End-to-end latency from input to structured output

Robustness Metrics.

- **Success Rate:** Percentage of documents processed without fatal errors
- **Per-Format Accuracy:** Extraction metrics broken down by document type (PNG, DOCX, etc.)

4.3 Results

We evaluate our multi-method extraction framework on a corpus of 400 synthetic Chinese identity documents spanning four formats (PNG, DOCX, XLSX, PDF), utilizing three extraction paradigms across 16 OCR-LLM configurations. Our findings underscore substantial variation in performance across both extraction strategies and document formats, with method efficacy closely aligned with the structural characteristics of each format.

Experimental Overview. The evaluation encompasses 16 extraction methods categorized into three distinct paradigms: (1) *Direct extraction*, which leverages document parsers or multimodal models; (2) *Replace-based extraction*, which employs rule-based template matching; and (3) *Table-based extraction*, which exploits spatial structure for field localization. Each method was evaluated on up to 100 documents per supported format, resulting in 2,500 test instances. We report precision, recall, F1 score, and processing latency, further decomposed into OCR and LLM inference time.

Table 1 summarizes the performance of the best and worst-performing methods for each document type. For structured formats (DOCX/XLSX), table-based methods achieve perfect F_1 scores (1.0), with docling_table and markitdown_table delivering 100% success rates while maintaining minimal latency (0.3–0.5s). In stark contrast, replace-based methods perform poorly on these formats, with markitdown_replace achieving only F_1=0.969 for both DOCX and XLSX, representing the worst performance with perfect rates dropping to 59% and 54% respectively. For image-based formats (PNG), the multimodal approach outperforms all OCR-based methods with an F_1 score of 0.999, while easyocr_table catastrophically fails with F_1=0.000 and 0% success rate despite minimal processing time (1.5s). PDF extraction exhibits the most extreme performance variance: docling_table maintains perfect accuracy (F_1=1.0) with low latency (1.6s), whereas mineru_replace completely fails (F_1=0.000, 0% success rate), highlighting severe method-format incompatibilities. Direct extraction methods consistently incur high LLM inference latency (13.4–13.6s), accounting for over 90% of

total processing time, while table-based approaches achieve 40–50$\times$ speedup through minimal LLM usage.

Figure 4 provides a holistic view of performance across all 16 methods and formats via dual heatmaps. The F_1 score heatmap reveals distinct performance patterns: table-based methods exhibit binary characteristics—either achieving near-perfect extraction ($F_1 \approx 1.0$) or complete failure ($F_1 = 0.0$), indicating strong format dependency. Direct extraction methods demonstrate more consistent but suboptimal performance across formats ($F_1 \approx 0.77$–0.99), while replace-based methods show the highest variability, ranging from moderate success to total failure. The processing time heatmap complements these findings, showing that methods with perfect F_1 scores often achieve the fastest processing times (0.3–1.6s for table-based approaches), while direct methods consistently require 13–15s due to LLM overhead. The multimodal method presents an outlier with 33.9s processing time for PNG files, trading computational efficiency for extraction accuracy. Empty cells in both heatmaps denote unsupported format-method combinations, particularly evident for specialized parsers like mineru (PDF-only) and OCR-based methods (image formats only). The visualization conclusively demonstrates that no single method achieves universal optimality across all formats, necessitating format-specific extraction strategies for optimal performance.

Image-based Documents (PNG). For scanned documents with identity cards, we observe stark contrasts between multimodal and OCR-based pipelines, as illustrated in Figure 5. The multimodal vision-language model demonstrates exceptional accuracy ($F_1 = 0.999 \pm 0.007$) through direct visual feature processing, effectively bypassing the cascading errors inherent in sequential OCR pipelines. However, this end-to-end approach incurs prohibitive computational costs, requiring 33.91 ± 9.49 seconds per document due to high-dimensional visual encoding and transformer-based attention mechanisms. Among OCR-based methods, we observe significant performance variations based on both the OCR engine and extraction strategy. PaddleOCR consistently outperforms EasyOCR across all extraction paradigms, achieving F_1 scores of 0.997, 0.961, and 0.997 for direct, replacement, and table-based extraction respectively, while EasyOCR exhibits substantially degraded performance ($F_1 = 0.760$, 0.688, and 0.000). This performance disparity stems fundamentally from EasyOCR's inability to preserve document spatial structure during text extraction, which disrupts the positional relationships critical for accurate field identification in structured documents like identity cards.

The spatial information loss particularly impacts table-based extraction, where EasyOCR completely fails ($F_1 = 0.000$) as the LLM cannot generate valid coordinates without reliable spatial encoding. In contrast, PaddleOCR maintains precise spatial mappings, enabling the table-based extraction strategy to achieve optimal performance at 0.63 ± 0.24 seconds—a $54\times$ speedup over the multimodal baseline—while preserving near-perfect accuracy ($F_1 = 0.997 \pm 0.011$). This method minimizes LLM inference to coordinate generation only, leveraging the preserved document structure to reduce computational overhead from $\sim$34 seconds to sub-second latency. The replacement method with PaddleOCR, employing regex patterns for standardized fields and batch processing for names, offers a middle ground at 0.72 ± 0.28 seconds ($F_1 = 0.961 \pm 0.044$), while direct extraction requires 13.75 ± 3.87 seconds despite high accuracy ($F_1 = 0.997 \pm 0.010$) due to processing entire OCR outputs. The PaddleOCR table-based approach thus emerges as the optimal solution for production deployment, combining superior spatial preservation with minimal LLM computation. The marginal 0.2% accuracy trade-off compared to multimodal methods represents an acceptable compromise given the 54-fold efficiency gain crucial for large-scale document processing systems.

Structured Office Documents (DOCX/XLSX). Our evaluation of structured office document extraction reveals fundamentally different performance characteristics compared to image-based formats, as illustrated in Figure 6. Native documents exhibit near-perfect accuracy across all extraction methodologies due to their inherent machine-readable structure and preserved spatial information. For DOCX files, both MarkItDown and Docling frameworks successfully extract the document's spatial structure, enabling high-fidelity extraction across all paradigms. Docling's table-based approach achieves optimal performance with perfect accuracy ($F_1 = 1.000$) and exceptional efficiency at 0.34 ± 0.03 seconds—a $41\times$ speedup compared to direct extraction (13.68 ± 3.85 seconds)—by leveraging structured table representations that minimize LLM processing to coordinate

Format	Method	Prec.	Rec.	F1	Succ. Rate	Perf. Rate	OCR (s)	LLM (s)	Total (s)
PNG	**multimodal**	**.999**	**.999**	**.999**	**100%**	**97%**	**—**	**—**	**33.9**
	paddleocr_table	.998	.997	.997	100%	93%	0.3	0.3	0.6
	paddleocr_direct	.998	.996	.997	100%	92%	0.4	13.4	13.8
	easyocr_table	**.000**	**.000**	**.000**	**0%**	**0%**	**1.2**	**0.3**	**1.5**
DOCX	**docling_table**	**1.00**	**1.00**	**1.00**	**100%**	**100%**	**0.1**	**0.3**	**0.3**
	markitdown_table	1.00	1.00	1.00	100%	100%	0.2	0.3	0.5
	docling_direct	1.00	1.00	1.00	100%	99%	0.1	13.6	13.7
	markitdown_direct	1.00	.999	1.00	100%	98%	0.2	13.6	13.8
	markitdown_replace	**.969**	**.969**	**.969**	**100%**	**59%**	**0.2**	**0.5**	**0.7**
XLSX	**markitdown_table**	**1.00**	**1.00**	**1.00**	**100%**	**100%**	**0.0**	**0.3**	**0.3**
	markitdown_direct	1.00	1.00	1.00	100%	99%	0.0	13.5	13.5
	markitdown_replace	**.969**	**.969**	**.969**	**100%**	**54%**	**0.0**	**0.5**	**0.5**
PDF	**docling_table**	**1.00**	**1.00**	**1.00**	**100%**	**100%**	**1.3**	**0.3**	**1.6**
	docling_direct	1.00	1.00	1.00	100%	99%	1.3	13.5	14.9
	mineru_direct	1.00	1.00	1.00	100%	99%	1.6	13.4	15.0
	mineru_replace	**.000**	**.000**	**.000**	**0%**	**0%**	**1.5**	**0.0**	**1.5**

Table 1: Comparative performance metrics showing best (green) and worst (red) extraction methods by document format. The multimodal method uses Qwen2.5-VL-7B while all other methods use Qwen2.5-7B. Prec. = Precision, Rec. = Recall, Succ. Rate = Success Rate (percentage of successful extractions), Perf. Rate = Perfect Rate (percentage achieving perfect F1 score of 1.0). OCR time represents character recognition overhead; LLM time indicates inference latency.

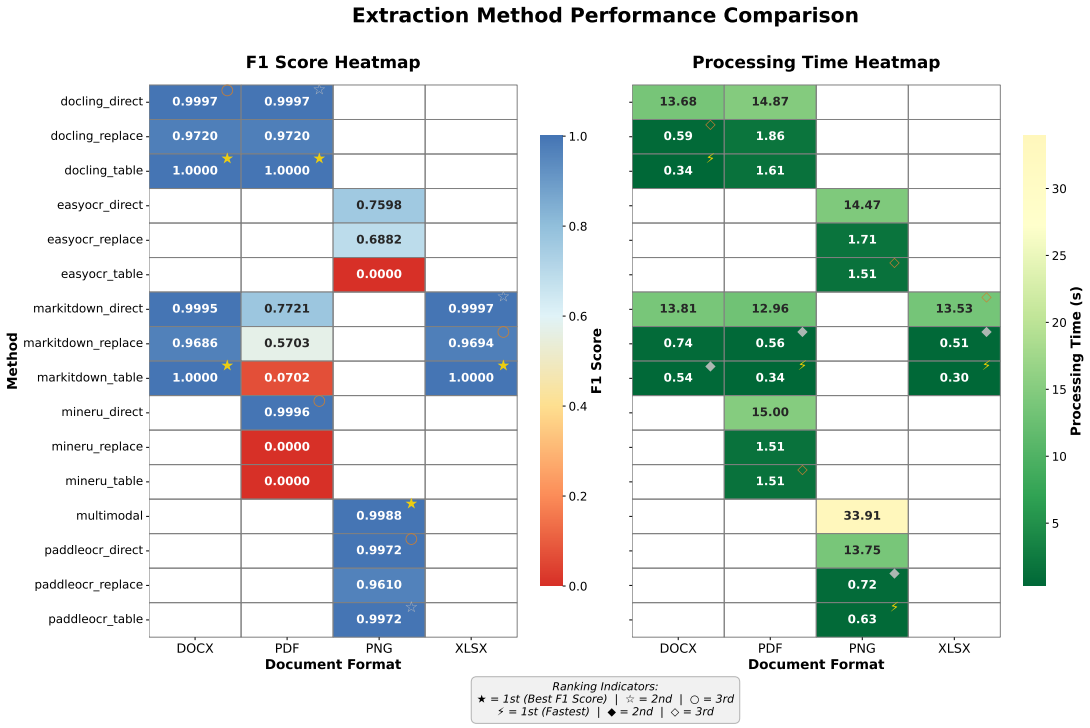

Figure 4: Performance comparison of extraction methods across document formats. The heatmap displays F_1 scores and processing time for 16 extraction methods (rows) tested on four document formats (columns). Empty cells denote unsupported format-method combinations. The multimodal method employs Qwen2.5-VL-7B, while all other methods utilize Qwen2.5-7B.

generation only. MarkItDown's table method follows closely with identical accuracy ($F_1 = 1.000$) at 0.54 ± 0.13 seconds, representing a $26\times$ improvement over its direct counterpart.

While replacement strategies offer competitive latency (Docling: 0.59 ± 0.13s, MarkItDown: 0.74 ± 0.18s), they sacrifice accuracy ($F_1 = 0.972 \pm 0.038$ and 0.969 ± 0.044 respectively) due to the LLM's occasional misidentification when matching ID numbers to corresponding names—a limitation stemming from the model's reliance on contextual inference rather than explicit structural

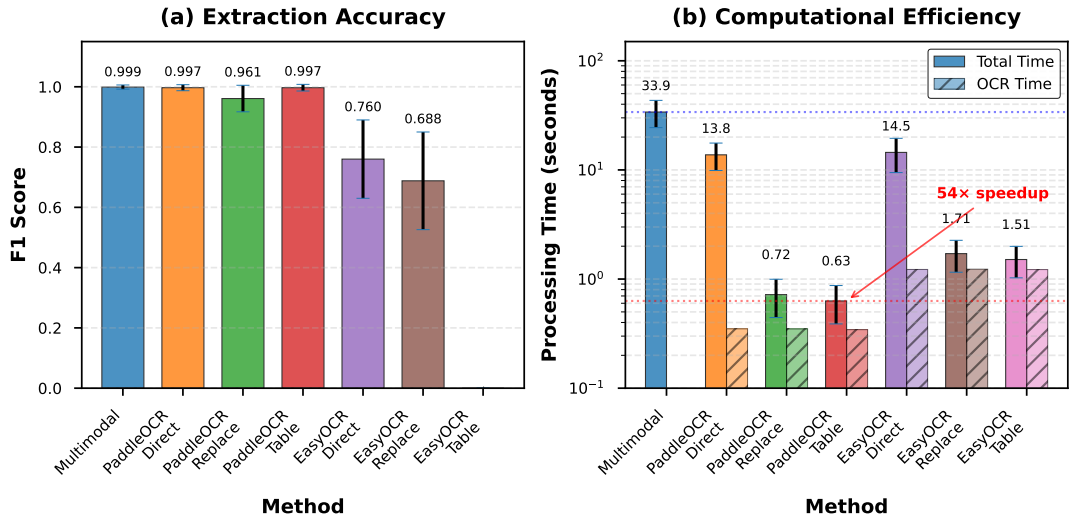

Figure 5: Performance on PNG-based documents: (a) F_1 score with standard deviation. (b) Latency comparison across pipelines.

cues. This name-matching error could be mitigated through prompt engineering to provide more explicit matching instructions or fine-tuning the general-purpose LLM on document-specific extraction tasks. For XLSX spreadsheets, where only MarkItDown provides support, the inherent tabular structure amplifies these efficiency gains: table-based extraction achieves perfect accuracy ($F_1 = 1.000$) with remarkable 0.30 ± 0.02 seconds processing time—a $44\times$ acceleration over direct methods (13.53 ± 3.82 seconds). This dramatic improvement stems from the natural alignment between spreadsheet cell structures and table-based extraction paradigms, eliminating the need for complex content interpretation. The replacement method maintains reasonable efficiency (0.51 ± 0.11s) but exhibits similar accuracy degradation ($F_1 = 0.969 \pm 0.039$) as observed in DOCX processing, with errors predominantly occurring in name-ID association tasks. These results establish table-based extraction as the unequivocally superior approach for native office documents, where the preserved document structure enables both perfect accuracy and minimal computational overhead, making it ideal for enterprise-scale document processing pipelines requiring both precision and throughput.

Portable Document Format (PDF). Our empirical analysis of PDF document processing reveals a fundamental trade-off between computational efficiency and extraction accuracy across different methodological approaches. MarkItDown prioritizes processing speed, achieving minimal OCR preprocessing overhead (0.056s), while Docling and MinerU adopt more computationally intensive strategies with OCR latencies of approximately 1.3–1.5s. However, this efficiency-accuracy trade-off manifests differently across extraction paradigms. MarkItDown's lightweight approach, which employs symbolic markers for spatial structure representation, exhibits severely degraded performance across all extraction strategies: direct extraction ($F_1 = 0.772$), replacement extraction ($F_1 = 0.570$), and table-based extraction ($F_1 = 0.070$).

The divergence in extraction efficacy between Docling and MinerU, despite their comparable preprocessing costs, underscores the critical importance of spatial representation strategies in document understanding. Docling maintains near-perfect accuracy across all extraction paradigms, with its table-based approach achieving optimal performance ($F_1 = 1.0$). In stark contrast, MinerU's reliance on HTML-style structural tags (e.g., `<td>`, `<tr>`) for encoding spatial relationships proves incompatible with text-based information extraction, resulting in catastrophic failure for replacement and table-based methods ($F_1 = 0.0$). The framework only maintains competitive performance in direct extraction mode ($F_1 = 0.9996$), where spatial structure parsing is bypassed entirely. These empirical findings, illustrated in Figure 7, establish Docling's table-based methodology as the optimal solution for PDF information extraction, successfully reconciling the competing demands of computational efficiency (1.61s total processing time)

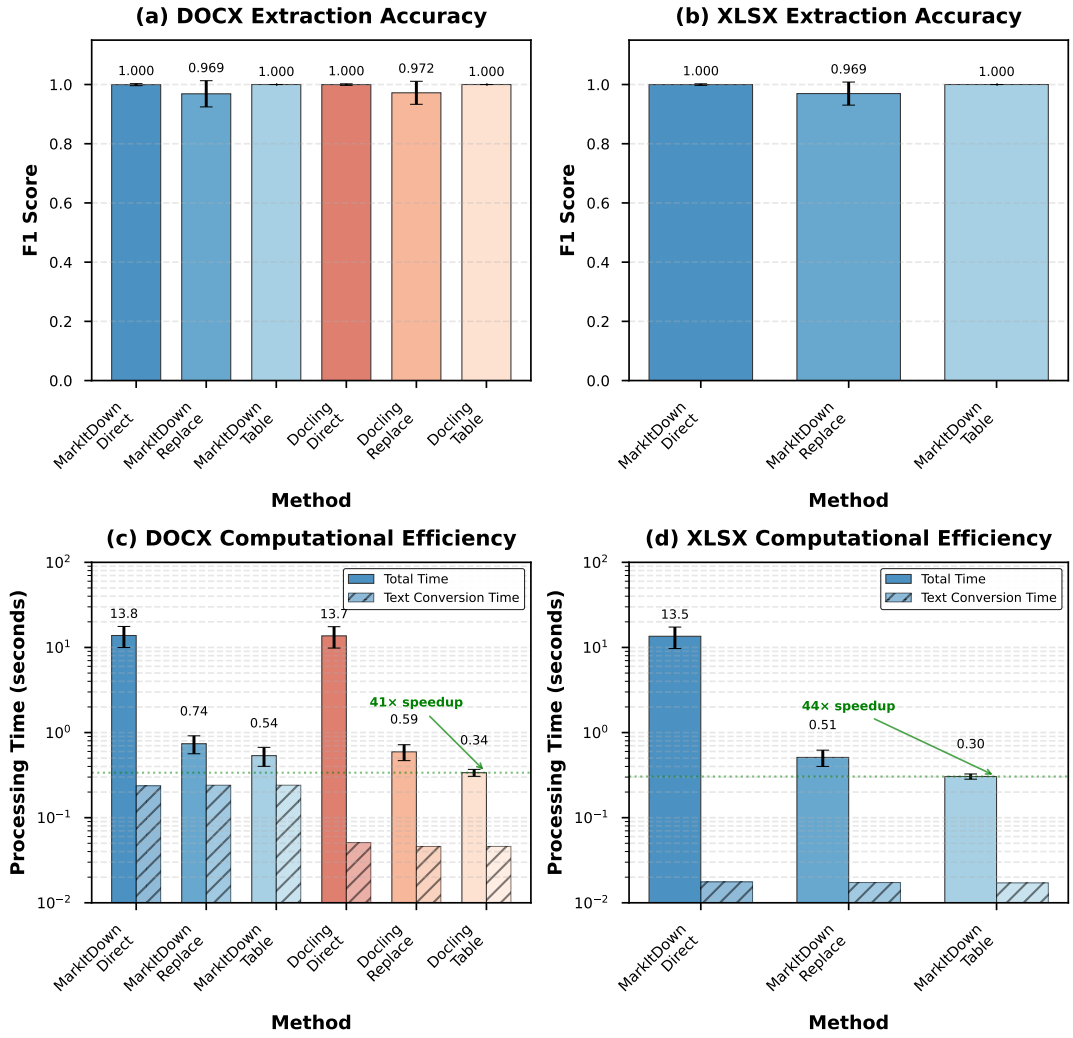

Figure 6: Performance comparison of different extraction methods for native office documents processing. (a) DOCX extraction accuracy. (b) XLSX extraction accuracy. (c) DOCX computational efficiency. (d) XLSX computational efficiency.

and extraction fidelity.

4.4 Discussion

Our comprehensive evaluation reveals a fundamental insight: the optimal extraction strategy for copy-heavy documents is intrinsically tied to document modality and structure, challenging the prevailing one-size-fits-all approaches in production systems. The 54× performance differential between table-based methods (0.97s) and multimodal approaches (33.91s) underscores a critical trade-off—while end-to-end models offer superior robustness through direct visual understanding, their computational overhead remains prohibitive for high-throughput scenarios characteristic of copy-heavy tasks. This finding suggests that the field's pursuit of universal extraction models may be misguided for repetitive document processing, where structural priors can be effectively exploited.

The stark performance dichotomy between OCR engines (PaddleOCR F1=0.997 vs. EasyOCR F1=0.000 for table extraction) reveals that spatial structure preservation, rather than character recognition accuracy alone, determines extraction success. This challenges conventional OCR evaluation metrics and highlights the need for structure-aware benchmarks. Moreover, the consistent superiority of table-based methods across native formats (100% accuracy at <1s) demonstrates that explicit structural modeling outperforms both semantic understanding (direct extraction) and pattern matching (replacement extraction) when document layouts are predictable—a defining characteristic of copy-heavy scenarios.

From a production standpoint, our results advo-

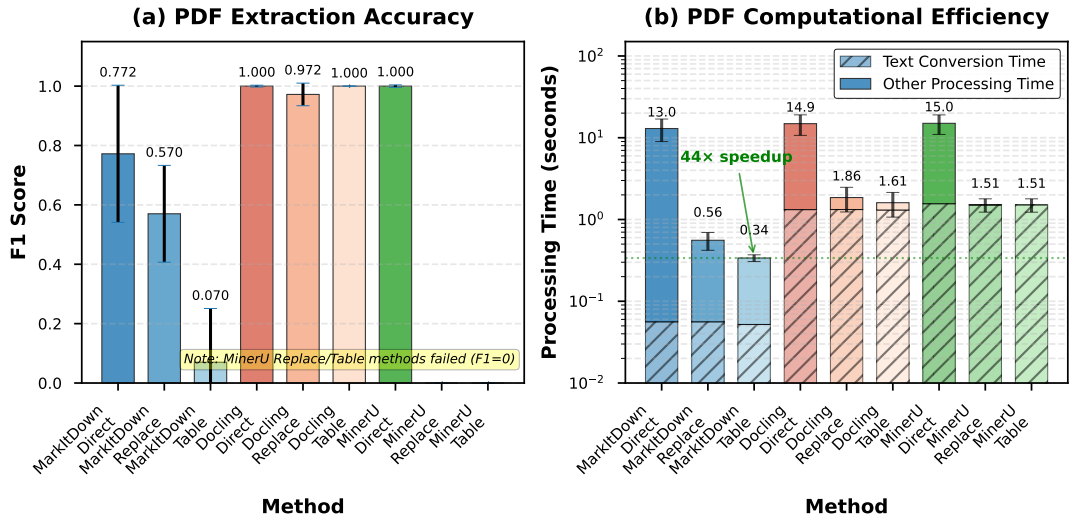

Figure 7: Performance comparison of different extraction methods for portable document format processing. (a) Extraction accuracy measured by F1 score with error bars showing standard deviation. (b) Computational efficiency comparison showing total latency (with error bars) and text conversion time for the markitdown, docling and minerU pipeline.

cate for adaptive, format-aware architectures over monolithic solutions. The multimodal approach's resilience to OCR failures positions it as an ideal fallback mechanism, while table-based extraction serves as the workhorse for structured documents. This hierarchical strategy—rapid format detection followed by method-specific routing—enables systems to process heterogeneous document streams at scale without sacrificing accuracy. The framework thus provides a blueprint for reconciling the competing demands of accuracy, efficiency, and robustness in real-world document processing pipelines.

5 Conclusion

We present a systematic framework for information extraction from copy-heavy documents, demonstrating that the repetitive nature of such tasks—rather than being a mere computational burden—can be strategically exploited through intelligent method selection. Our comprehensive evaluation of 25 method combinations across diverse documents reveals that optimal extraction strategies must align with document characteristics: table-based methods for structured formats (achieving perfect accuracy at sub-second latency), multimodal approaches for degraded images (F1=0.999), and adaptive routing for heterogeneous streams.

The core contribution lies not in proposing novel extraction techniques, but in establishing a principled approach to method selection for copy-heavy scenarios. By recognizing that document structure dictates optimal extraction paradigms, we achieve a 54× speedup while maintaining accuracy—a critical requirement for enterprise-scale deployment. This work challenges the field's emphasis on universal models, showing that domain-specific solutions remain indispensable when processing millions of similar documents daily.

For practitioners, our framework offers immediate value: implement table-based extraction as the default for structured documents, maintain multimodal models as robust fallbacks, and use format detection for intelligent routing. For researchers, we highlight unexplored opportunities in hybrid architectures that combine the efficiency of structural methods with the robustness of end-to-end models. As document processing increasingly underpins digital transformation, our work provides both theoretical insights and practical tools for building extraction systems that scale without compromising quality.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Peter G Anick and Rex A Flynn. 1992. Versioning a full-text information retrieval system. In *Proceedings of the 15th annual international ACM SIGIR confer-*

ence on Research and development in information retrieval, pages 98–111.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 993–1003.

Douglas E Appelt. 1999. Introduction to information extraction. *Ai Communications*, 12(3):161–172.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model apis. *arXiv preprint arXiv:2301.08721*.

Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832.

Dustin Cobas and Gonzalo Navarro. 2019. Fast, small, and simple document listing on repetitive text collections. In *International Symposium on String Processing and Information Retrieval*, pages 482–498. Springer.

Ciaran Cooney, Joana Cavadas, Liam Madigan, Bradley Savage, Rachel Heyburn, and Mairead O'Cuinn. 2023. End-to-end document classification and key information extraction using assignment optimization. *arXiv preprint arXiv:2306.00750*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 363–370.

Dayne Freitag. 2000. Machine learning for information extraction in informal domains. *Machine learning*, 39(2):169–202.

Travis Gagie, Aleksi Hartikainen, Kalle Karhu, Juha Kärkkäinen, Gonzalo Navarro, Simon J Puglisi, and Jouni Sirén. 2017. Document retrieval on repetitive string collections. *Information Retrieval Journal*, 20:253–291.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. 2024. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338.

Beliz Gunel, Navneet Potti, Sandeep Tata, James B Wendt, Marc Najork, and Jing Xie. 2022. Data-efficient information extraction from form-like documents. *arXiv preprint arXiv:2201.02647*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7(15):2.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. *Preprint*, arXiv:2203.08411.

J Li, J Chen, R Ren, X Cheng, WX Zhao, JY Nie, and JR Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. arxiv, article. *arXiv preprint arXiv:2401.03205*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for zero-shot open-domain qa. *arXiv preprint arXiv:2212.08635*.

Yiming Lin, Mawil Hasan, Rohan Kosalge, Alvin Cheung, and Aditya G Parameswaran. 2025. Twix: Automatically reconstructing structured data from templatized documents. *arXiv preprint arXiv:2501.06659*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 6495–6504.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Gonzalo Navarro. 2019. Document listing on repetitive collections with guaranteed performance. *Theoretical Computer Science*, 772:58–72.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, and 1 others. 2023. Lmdx: Language model-based document information extraction and localization. *arXiv preprint arXiv:2309.10952*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Filip Seitl, Tomáš Kovářík, Soheyla Mirshahi, Jan Kryštůfek, Rastislav Dujava, Matúš Ondreička, Herbert Ullrich, and Petr Gronat. 2024. Assessing the quality of information extraction. *arXiv preprint arXiv:2404.04068*.

Xiaoyu Shen, Youssef Oualil, Clayton Greenberg, Mittul Singh, and Dietrich Klakow. 2017. Estimation of gap between current language models and human performance. In *Proc. Interspeech 2017*, pages 553–557.

Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. 2022. Rocbert: Robust chinese bert with multimodal contrastive pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–931.

Liyan Tang, Dhruv Rajan, Suyash Mohan, Abhijeet Pradhan, R Nick Bryan, and Greg Durrett. 2021. Making document-level information extraction right for the right reasons. *arXiv preprint arXiv:2110.07686*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023a. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Liqiang Wang, Xiaoyu Shen, Gerard de Melo, and Gerhard Weikum. 2020. Cross-domain learning for classifying propaganda in online contents. *arXiv preprint arXiv:2011.06844*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, and 1 others. 2023c. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, and 1 others. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv e-prints*, pages arXiv–2302.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Conghui He, and Wentao Zhang. 2024. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023a. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, and 1 others. 2023b. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710.

Zhen Zheng, Xin Ji, Taosong Fang, Fanghao Zhou, Chuanjie Liu, and Gang Peng. 2024. Batchllm: Optimizing large batched llm inference with global prefix sharing and throughput-oriented token batching. *arXiv preprint arXiv:2412.03594*.