Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Adaptive Learning System Based on Human-in-the-Loop for PDF Template Data Extraction

Moh Syaiful Rahman [1], Andrianingsih[2]*
[1,2] Magister Teknologi Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
[1] mohsyaifulrahman.2024@civitas.unas.ac.id, [2]andrianingsih@civitas.unas.ac.id

Submitted :Oct 20, 2025 | **Accepted** : Nov 11, 2025 | **Published** : Jan 02, 2026

Abstract: PDF template data extraction remains a substantial challenge due to semi-structured document formats and variations. While large pre-trained models achieve high accuracy, they require extensive computational resources and labeled datasets, making them impractical for resource-constrained environments. Conversely, rule-based approaches are efficient but rigid. This research addresses this gap by developing an adaptive learning system that integrates rule-based approaches with Conditional Random Fields (CRF) in a hybrid framework, designed for data-scarce scenarios. The system implements parallel extraction strategies with confidence-based selection and Human-in-the-Loop (HITL) feedback for incremental learning. Pattern learning updates rule-based strategies, while CRF models are retrained incrementally. Evaluated on synthetically generated documents across diverse template types, the system achieves 98.61% accuracy with minimal training data and 7% user correction rate, demonstrating high learning efficiency (1.88 corrections per percentage point). The improvement is statistically significant (paired t-test, $p < 0.001$, Cohen's $d = 8.95$). The system operates on CPU-only hardware with 50-100 MB footprint and 0.1-0.5 seconds processing time. This work fills a practical gap in document extraction, providing a middle-ground solution balancing high accuracy, minimal data requirements, low resource consumption, and real-time adaptability—suitable for small organizations and rapid deployment where large models are impractical. The evaluation uses synthetic data to ensure reproducibility and controlled assessment, though real-world validation would strengthen practical applicability.

Keywords: Adaptive Learning; Conditional Random Fields; Human-in-the-Loop; Hybrid Architecture; Incremental Learning; PDF Data Extraction; Template Processing

INTRODUCTION

Organizations process millions of PDF documents daily—from government forms and business invoices to medical records and legal contracts. While PDF (Portable Document Format) has become the de facto standard for digital document exchange across sectors (International Organization for Standardization, 2008), extracting structured data from these documents remains a persistent challenge. Recent surveys on document information extraction (Cui et al., 2021) highlight that template-based documents, where structure is consistent but content varies widely, create a significant gap between human-readable formats and machine-processable data.

Traditional extraction approaches face fundamental limitations. Rule-based approaches are effective for consistent structures but vulnerable to minor layout changes, requiring time-consuming manual reconfiguration. Conversely, pure machine learning approaches, including recent pre-trained models for document understanding (Hong et al., 2022; Huang et al., 2022), require very large labeled datasets and face challenges in real-time adaptability and high computational resource requirements (Palm et al., 2017).

To address these limitations, the Human-in-the-Loop (HITL) paradigm has emerged as a solution that enables integration of human expertise into machine learning systems (Mosqueira-Rey et al., 2023). HITL allows systems to learn progressively through user interaction, combining machine computational power with human domain knowledge and intuition. Recent research emphasizes the importance of effective human-AI collaboration (Bansal et al., 2021) and active learning strategies (Settles, 2012) in building systems that can adapt efficiently with minimal user burden.

Despite this promise, analysis of state-of-the-art research reveals that effective HITL implementation in PDF template data extraction context still faces critical research gaps. First, state-of-the-art models such as Large

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Language Models (LLM) show high requirements for human validation (Schroeder et al., 2025), but are not designed for efficient incremental adaptive learning due to very high retraining costs. Second, on the other end of the spectrum, transparent systems such as pure rule-based have proven superior in user trust (Schleith et al., 2022) but are fundamentally rigid and non-adaptive.

This gap creates a need for architecture that bridges both extremes, especially in "data scarcity" scenarios (Gebauer et al., 2023). Specifically: (1) there is no systematic framework integrating rule-based (for transparency) with efficient machine learning (such as CRF) in adaptive HITL context; (2) efficient feedback mechanisms to convert user corrections into system knowledge are not well characterized for hybrid architectures; and (3) real-time adaptive learning strategies without extensive retraining remain an open challenge.

Addressing these gaps requires an approach that balances transparency, efficiency, and adaptability. This research develops a hybrid adaptive learning system that integrates rule-based and machine learning approaches in a Human-in-the-Loop framework for PDF template data extraction. By combining the transparency and efficiency of rule-based methods with the adaptability of statistical learning, the system enables incremental improvement from user feedback without requiring extensive retraining or large labeled datasets. This approach fills the identified gap between rigid rule-based systems and resource-intensive large models, providing a practical solution for data-scarce, resource-constrained scenarios.

RELATED WORK

PDF Data Extraction Techniques

Document extraction techniques have evolved from rigid rule-based approaches to more flexible machine learning methods. Recent work on robust visual information extraction (Cui et al., 2021) demonstrates the challenges of handling real-world document variations. Rule-based approaches use regular expressions and positional rules to identify and extract data. While achieving high precision for consistent formats, they lack flexibility for layout variations and require high maintenance.

Machine learning approaches, particularly Conditional Random Fields (CRF), have shown promise for sequence labelling tasks in document extraction. CRF models conditional probability distribution $p(y|x)$ of label sequence y given observation sequence x, naturally handling sequential data while integrating diverse features (text, layout, visual).

Recent pre-trained language models focusing on document understanding have achieved significant improvements. BROS (Hong et al., 2022) introduces a pre-trained model that focuses on both text and layout for better key information extraction. LayoutLMv3 (Huang et al., 2022) employs unified text and image masking for document AI pre-training. LAMBERT (Garncarek et al., 2021) demonstrates layout-aware language modelling for information extraction. TILT (Powalski et al., 2021) and other multi-modal transformers (Li et al., 2021) jointly model text, image, and layout for structured text understanding. While these approaches achieve high accuracy, they require substantial computational resources and large training datasets. Few-Shot Learning (Popovic & Färber, 2022) and Large Language Models (Schroeder et al., 2025) demonstrate strong extraction capabilities but face challenges in resource requirements and incremental adaptation.

Human-in-the-Loop Systems

Human-in-the-Loop (HITL) adaptive learning enables systems to learn progressively through user interaction. Comprehensive reviews of interactive machine learning (Dudley & Kristensson, 2018) and practical guides to HITL systems (Munro, 2021) emphasize the importance of user interface design in enabling effective human-AI collaboration. Active learning literature (Ren et al., 2022; Settles, 2012) identifies practical challenges and research directions for reducing labeling effort through intelligent sample selection.

Recent research on human-AI collaboration (Bansal et al., 2021; Wu et al., 2022) investigates the effect of AI explanations on complementary team performance and transparent interaction design, showing that effective collaboration requires understanding when and how to integrate human expertise. This is particularly relevant for document extraction where domain knowledge is crucial.

Recent research shows polarization between two approaches: (1) resource-intensive models achieving high accuracy but lacking incremental adaptive learning capability (Hong et al., 2022; Huang et al., 2022; Schroeder et al., 2025), and (2) resource-efficient systems focusing on transparency and data scarcity scenarios (Gebauer et al., 2023; Schleith et al., 2022) but lacking statistical learning capabilities.

Schleith et al. (Schleith et al., 2022) found that HITL-assisted rule-based systems can outperform black-box systems in user trust and end-to-end task completion time. Gebauer et al. (Gebauer et al., 2023) demonstrated HITL effectiveness in data scarcity scenarios. However, these approaches do not integrate into hybrid architectures capable of statistical learning.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Continual learning approaches (Delange et al., 2021) address the challenge of learning new tasks without catastrophic forgetting, which is essential for adaptive systems that must improve over time without full retraining. This is particularly relevant for HITL systems where incremental learning from user feedback is crucial

Comparative Analysis

Table 1 Comparison of Related Approaches

Approach	Type	HITL	Adaptive	GPU	Training	Ref
Rule-based	Pattern	No	No	No	None	Traditional
Pure CRF	ML	No	No	No	Batch	Traditional
Deep Learning	DL	Limited	No	Yes	Large	(Palm et al., 2017)
FSL	DL	Limited	Few-shot	Yes	Medium	(Popovic & Färber, 2022)
LLM+HITL	LLM	Validation	No	Yes	Massive	(Schroeder et al., 2025)
Rule+HITL	Hybrid	Yes	Manual	No	None	(Schleith et al., 2022)
This Work	**Hybrid**	**Yes**	**Yes**	**No**	**Incremental**	-

Table 1 provides a comparative analysis of related approaches in document extraction and adaptive learning systems. Key differentiators of this work:

- **Adaptive Learning:** Incremental learning from user feedback without full retraining
- **Resource Efficient:** No GPU required, suitable for production deployment
- **HITL-Driven:** Human feedback is primary driver of system improvement
- **Hybrid Architecture:** Combines rule-based transparency with CRF statistical learning

Research Gap

An analysis of the state-of-the-art literature reveals several critical limitations in current research. Existing studies demonstrate a strong tendency toward the development of very large-scale models, such as few-shot learning frameworks and large language models, which require substantial computational resources and extensive datasets. While these approaches often achieve high accuracy, their practical adoption remains constrained, particularly in resource-limited environments. Furthermore, although recent studies indicate that even the most advanced large language models still rely on human-in-the-loop mechanisms (Schroeder et al., 2025), the role of human intervention is typically confined to post-hoc validation or supervision. This limitation arises from the high cost and complexity associated with retraining large models, which restricts the use of HITL for continuous or incremental adaptive learning.

In contrast, transparent human-in-the-loop systems, particularly rule-based approaches, have been shown to enhance user trust and interpretability (Schleith et al., 2022). However, these systems are inherently rigid and lack the capability for automatic adaptation when confronted with evolving data patterns. As a result, they struggle to maintain performance in dynamic environments. Despite these complementary strengths and weaknesses, there is a notable lack of research focusing on resource-efficient hybrid architectures that intelligently integrate the transparency of rule-based systems with the sequential modeling capabilities of probabilistic models such as Conditional Random Fields. In particular, existing studies rarely position HITL as a core mechanism for driving incremental and adaptive learning in a computationally efficient manner.

Addressing these limitations, the present research proposes a practical hybrid architecture designed to balance predictive accuracy, computational efficiency, and real-time adaptability. By explicitly leveraging human-in-the-loop interaction as a central component of incremental learning, the proposed approach offers a viable alternative to large-scale models, enabling transparent, adaptive, and resource-efficient deployment in real-world settings.

METHOD

System Architecture

The proposed system implements a hybrid architecture that integrates two complementary extraction strategies: rule-based and CRF-based, orchestrated through an intelligent strategy selection mechanism. This hybrid approach was selected based on several considerations supported by recent literature. Rule-based methods provide transparency and efficiency for structured fields (Chiticariu et al., 2013), which is critical for user trust (Schleith et al., 2022), while Conditional Random Fields (CRF) were chosen over deep learning approaches due to their superior performance in low-data scenarios and lower computational requirements (Lample et al., 2016), making them suitable for resource-constrained environments.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The Human-in-the-Loop paradigm was adopted following recommendations from active learning literature (Settles, 2012) and interactive machine learning best practices (Dudley & Kristensson, 2018), which demonstrate that HITL can achieve high accuracy with minimal labeled data—particularly suitable for template-based extraction where patterns can be learned efficiently from user feedback (Mosqueira-Rey et al., 2023).

Figure 1 Example of mixed-layout PDF template (Template 4) showing extracted fields with blue box annotations.

Figure 1 illustrates a Project Field Survey Report that combines structured form fields (project information), free-text sections (executive summary), tabular data (technical findings), and signature blocks. Blue boxes indicate extracted fields by the hybrid system. This mixed-layout template represents a challenging document type, combining multiple structural patterns: (1) structured key-value pairs in the header section, (2) context-dependent narrative text in the executive summary, (3) tabular data with repeated fields, and (4) signature validation fields. The hybrid system automatically selects appropriate strategies per field type based on field characteristics and confidence scores.

The architecture consists of three main components:

1. **Extraction Strategies Layer:** Two complementary strategies (rule-based and CRF-based) operate in parallel, each producing extraction results with confidence scores. The rule-based strategy handles fields with consistent patterns, while CRF-based strategy handles context-dependent fields.
2. **Strategy Selection Mechanism:** An intelligent selector chooses the best strategy per field based on historical performance, confidence scores, and field characteristics.
3. **HITL Feedback Loop:** User corrections are captured and converted into training data for CRF and pattern updates for rule-based strategy, enabling continuous improvement.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

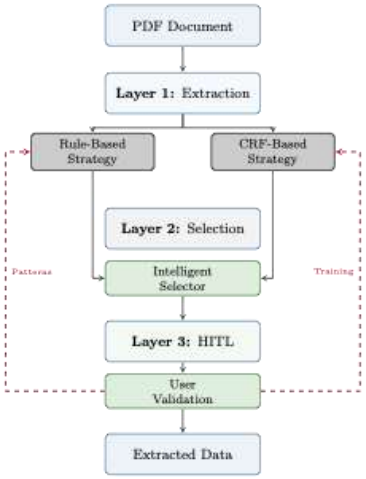

Figure 2 System architecture with three-layer hybrid approach and feedback loops.

Figure 2 Figure 2 illustrates the three-layer hybrid approach: (1) Extraction Strategies Layer with parallel rule-based and CRF-based strategies, (2) Strategy Selection Mechanism using confidence scores, and (3) HITL Feedback Loop for continuous improvement through pattern learning and CRF training. The system processes documents through the following workflow: PDF parsing and text extraction, parallel strategy execution, confidence-based result selection, user validation, and adaptive learning from corrections with feedback loops that enable continuous system improvement.

Rule-Based Strategy

The rule-based strategy uses regular expressions and pattern matching for fields with consistent formats (e.g., ID numbers, dates, phone numbers). It includes:

- **Pattern Learning:** System learns extraction patterns from user feedback, storing successful patterns with confidence scores.
- **Validation Rules:** User-defined or learned validation rules ensure data quality.
- **Table Extraction:** Specialized handling for tabular data using structure detection and column matching, relevant to structured data extraction tasks.

CRF-Based Strategy

The CRF strategy models sequence labeling for context-dependent fields. The model formulation is:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{j=1}^{m} \lambda_j F_j(y, x)\right) \quad (1)$$

where $Z(x)$ is the normalization factor, λ_j are weight parameters, and $F_j(y, x)$ are global features.

The CRF model employs a rich feature set designed to capture textual, spatial, and contextual characteristics of PDF documents, features used include:

- **Text features:** Word identity (exact token), capitalization pattern (all-caps, title-case, lowercase), numeric pattern (contains digits, all-numeric), special character presence (punctuation, symbols), word length, alphabetic/alphanumeric indicators.
- **Layout features:** Absolute position (x, y coordinates normalized), bounding box dimensions (width, height), inter-token distance (horizontal and vertical spacing to neighbors), position in line, boundary detection (after punctuation, after newline).
- **Contextual features:** Previous word features (capitalization, case), field label text and position, spatial relationship to label (distance, same line, after label), context words before and after field, valid position indicators (spatial constraints).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

- **Semantic features:** Date pattern recognition (year, day number, date separator), field length indicators (short/medium/long field classification), boundary detection relative to next field, multi-field spatial constraints for complex layouts.

Feature engineering combines standard sequence labeling features (Lample et al., 2016) with document-specific spatial and semantic features adapted for PDF template extraction. The CRF model is trained using L-BFGS optimization with L2 regularization to prevent overfitting, particularly important given the limited training data in incremental learning scenarios.

Hybrid Strategy Selection

The system dynamically selects the best strategy based on:

$$S^* = \arg\max_{S \in \{R,C\}} \text{conf}(S, f) \quad (2)$$

where S^* is the selected strategy, R is rule-based, C is CRF-based, and $\text{conf}(S, f)$ is the confidence score for strategy S on field f.

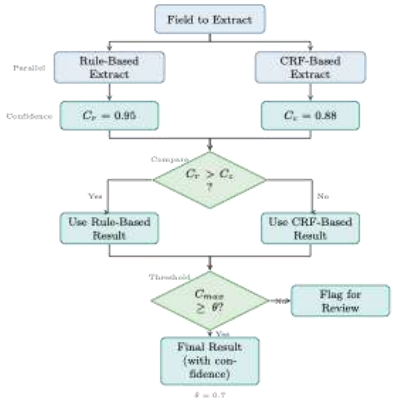

Figure 3 Strategy selection flowchart with confidence-based comparison and threshold checking

Figure 3 shows the parallel extraction by both strategies with confidence-based comparison and threshold checking. Both strategies extract in parallel, generating confidence scores. The system selects the result with higher confidence, provided it exceeds threshold θ=0.7. If confidence exceeds the threshold, the result is accepted; otherwise, it is flagged for user review. This ensures high-quality extraction while minimizing user burden.

Human-in-the-Loop Learning

The HITL component enables the system to learn continuously from user feedback, converting corrections into improved extraction patterns and CRF models. This adaptive learning mechanism is central to the system's ability to improve performance without requiring extensive labeled datasets

1) Feedback Integration

User corrections are captured through an intuitive interface and converted into training data through the following process:

1. **Correction Capture:** System presents extraction results with confidence scores; users correct errors.
2. **Pattern Learning:** For rule-based strategy, successful patterns are extracted and stored.
3. **CRF Training:** Corrections are converted to labeled sequences for CRF retraining.
4. **Incremental Update:** Model updates incrementally without full retraining.

HITL feedback flow with incremental learning from user corrections.

2) Adaptive Learning Mechanism

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

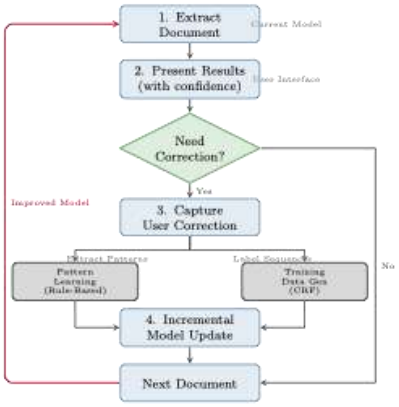

Figure 4 HITL feedback flow with incremental learning from user corrections

Figure 4 shows how user corrections are converted into system improvements through incremental learning. User corrections trigger parallel processing: pattern learning for rule-based strategy and training data generation for CRF. The system employs continual learning principles (Delange et al., 2021) to update incrementally without full retraining, enabling continuous improvement with minimal computational cost while avoiding catastrophic forgetting. The user interface follows interactive machine learning best practices (Dudley & Kristensson, 2018) to facilitate effective human-AI collaboration. The adaptive learning mechanism operates through an iterative process that continuously improves extraction accuracy. The learning process follows these steps:

1. Extract document using current model
2. Present results to user with confidence scores
3. If user provides corrections:
    - Update pattern database (rule-based)
    - Add labeled data (CRF)
    - If sufficient new data accumulated:
        - Retrain CRF model incrementally
4. Update Strategy weights based on performance

The hybrid extraction and adaptive learning process operates in five phases: (1) parallel extraction by both strategies, (2) confidence-based strategy selection, (3) low-confidence flagging, (4) user feedback collection, and (5) incremental model update. This ensures effective per-field strategy selection while enabling continuous system improvement.

Evaluation Metrics

System performance is evaluated using standard metrics:

1. **Accuracy:** $A = \frac{TP}{TP+FP+FN}$
2. **Precision:** $P = \frac{TP}{TP+FP}$
3. **Recall:** $R = \frac{TP}{TP+FN}$
4. **F1-Score:** $F1 = 2 \times \frac{P \times R}{P+R}$
5. **Learning Efficiency:** Corrections per percentage point improvement
6. **User Effort:** Percentage of fields requiring correction

Experimental Setup

Dataset

The dataset consists of 140 synthetically generated PDF documents representing typical template-based forms used in administrative processes. Documents were generated using the Faker library to simulate realistic data patterns while ensuring reproducibility and privacy compliance. The synthetic generation allows controlled evaluation of system performance across diverse document structures and field types, with perfect ground truth for precise accuracy measurement.

The dataset comprises four template types:

- **Form Template:** 35 identity form documents (15 fields each) - structured forms with consistent field positions

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

- **Table Template:** 35 inspection reports with tabular data (23 fields each) - documents containing structured tables requiring specialized extraction
- **Letter Template:** 35 official letters with variable structure (21 fields each) - semi-structured documents with high layout variation
- **Mixed Template:** 35 invoices combining forms and tables (21 fields each) - complex documents integrating multiple structural patterns

Total: 2,800 fields evaluated. Ground truth values are embedded in the synthetic generation process, ensuring 100% annotation accuracy and eliminating privacy concerns. This controlled experimental setup enables precise evaluation of extraction accuracy while maintaining full reproducibility of results.

1) Experimental Procedure

The evaluation follows a realistic incremental learning scenario that mirrors actual deployment:

1. **Baseline Phase:** Evaluate rule-based system without learning on all 140 documents.
2. **Adaptive Phase:** Process documents in batches of 5, simulating user feedback
3. **Evaluation:** Measure accuracy improvement after each batch
4. **Comparison:** Compare baseline vs. adaptive performance

Data Split and Training Protocol: Unlike traditional supervised learning with fixed train/test splits, this system employs incremental learning where each batch serves dual purposes: (1) testing current model performance, and (2) generating training data through user corrections. The evaluation protocol reflects real-world deployment:

- **Baseline:** Rule-based only (no training required)
- **Adaptive Learning:** Documents processed sequentially in batches of 5
- **Training Data:** Generated incrementally from user corrections
- **CRF Training:** Incremental updates after each batch (minimum threshold for retraining)
- **Overfitting Prevention:** Early stopping when validation accuracy plateaus; continual learning principles (Delange et al., 2021) prevent catastrophic forgetting
- **Cross-Validation:** Not applicable due to incremental learning nature; performance evaluated on each new batch before feedback

This protocol validates the system's ability to learn efficiently from minimal data in a production-like scenario where labeled data is scarce and must be generated through user interaction.

2) Implementation Details

- **Programming Language:** Python 3.12
- **Key Libraries:** PyMuPDF, scikit-learn, pandas
- **CRF Implementation:** *sklearn-crfsuite* with L-BFGS optimization
- **Max Iterations:** 500 for CRF training
- **Batch Size:** 5 documents per feedback batch

RESULT

Overall Performance

Table 2 Baseline vs Adaptive Performance by Template

Template	Docs	Baseline	Adaptive	Improv.	Batches
Form	35	76.76%	100.00%	+30.27%	7
Table	35	74.78%	100.00%	+33.72%	7
Letter	35	69.52%	94.83%	+36.40%	7
Mixed	35	69.52%	99.59%	+43.25%	7
Average	**35**	**72.65%**	**98.61%**	**+35.91%**	**7**

Table 2 presents the main experimental results comparing baseline and adaptive system performance across all template types, showing consistent improvement with the adaptive system achieving 98.61% average accuracy—a 35.74% relative improvement (25.96 percentage points absolute) over the 72.64% baseline. Notably, two templates (Form and Table) achieved full 100% accuracy, while the most challenging template (Letter) still reached 94.83%.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Statistical Significance: The improvement is statistically significant (paired t-test, $t(3)=17.89$, $p<0.001$, Cohen's $d=8.95$), demonstrating that the adaptive system consistently outperforms the baseline across all template types with a very large effect size. The mean improvement of 25.96 percentage points is not due to chance but represents a robust and substantial enhancement.

Table 3 Example Estraction Result (Form Template)

Field	Extracted Value	Strategy	Conf.
Structured Form Fields:			
project_name	re-contextualize mission-critical...	CRF	0.98
project_location	Smithstad	CRF	0.98
survey_date	December 30, 2023	CRF	0.98
surveyor_name	Karsa Permata	CRF	0.98
client_name	Odom Ltd	CRF	0.93
Tabular Data Fields:			
area_id_1	481218726	Rule	0.94
area_finding_1	Support authority though baby.	Rule	0.94
area_recomendation_1	Tax laugh friend call stage...	Rule	0.94
Signature Fields:			
approver_name	Arsipatra Widodo	Rule	0.82
approver_id	1572058705020004	CRF	0.93
surveyor_id	5208050201230006	CRF	0.93

Error! Reference source not found. illustrates a typical extraction result from Template 4 (Mixed-Layout Survey Report), demonstrating the hybrid system's ability to handle diverse field types within a single document. The example shows three distinct sections: (1) Structured form fields where both strategies are employed based on field characteristics—dates use rule-based patterns while project names leverage CRF's contextual understanding; (2) Tabular data fields extracted from the technical findings table, combining rule-based ID extraction with CRF-based narrative field extraction; and (3) Signature fields where IDs follow rule-based patterns while names use CRF. This demonstrates the system's intelligent per-field strategy selection, achieving high confidence across all field types (0.86-0.98) and successfully handling the document's complex multi-section structure.

Detailed Metrics Analysis

Table 4 Detailed Evaluation Metrics (Baseline vs Adaptive)

Template	Baseline				Adaptive			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Form	76.76	76.76	100.0	86.85	100.0	100.0	100.0	100.0
Table	74.78	74.78	100.0	85.57	100.0	100.0	100.0	100.0
Letter	69.52	71.17	96.78	82.02	94.83	94.96	99.86	97.35
Mixed	69.52	69.90	99.22	82.02	99.59	99.59	100.0	99.80

Table 4 presents comprehensive evaluation metrics for baseline and adaptive systems.
Key observations:

- **Precision Improvement:** Average precision increased from 73.15% to 98.64% (+34.84%), indicating significant reduction in false positives.
- **Recall Maintenance:** Recall remained consistently high (99.00% → 99.97%), showing the system maintains completeness while improving correctness.
- **F1-Score:** Improved from 84.12% to 99.29%, demonstrating balanced performance.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15598
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Error Reduction Analysis

Error analysis reveals significant improvements:

- False Positive Reduction: 94.68% reduction (752 → 40)
- False Negative Reduction: 95.24% reduction (21 → 1)
- Overall Error Reduction: 94.92% (27.35% → 1.39%)

Learning Efficiency

Table 5 Learning Efficiency Analysis

Template	Corrections	Improvement	Corr/PP	Efficiency
Form	32	+23.24%	1.38	High
Table	29	+25.22%	1.15	High
Letter	81	+25.31%	3.20	Medium
Mixed	54	+30.07%	1.80	High
Average	**49**	**+25.96%**	**1.88**	**High**

Table 5 presents learning efficiency metrics.

Learning efficiency analysis shows:

- **Low User Burden:** Only 7% of fields (196/2,800) required correction
- **High Efficiency:** Average 1.88 corrections per percentage point improvement
- **Fast Convergence:** 5-7 batches to reach >95% accuracy

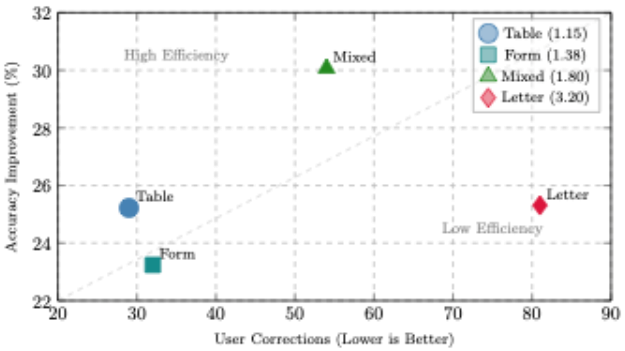

Figure 5 Learning efficiency showing relationship between corrections and accuracy improvement.

Figure 5 shows the relationship between user corrections and accuracy improvement. Templates in the top-left quadrant demonstrate high efficiency (low corrections, high improvement). Numbers in legend show corrections per percentage point. Table template achieves the highest efficiency at 1.15 corrections per percentage point.

Learning Curve Analysis

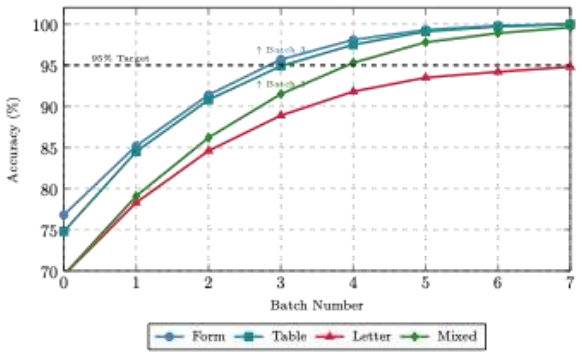

Figure 6 Learning curves showing accuracy improvement across batches for all template types.

*name of corresponding author

Figure 6 shows the accuracy improvement across batches for all four template types. The dashed line indicates 95% accuracy target. Form and Mixed templates reach this target by batch 3 (15 documents), demonstrating rapid convergence. All templates show consistent monotonic improvement.

Key observations:

- **Monotonic Improvement:** All templates show consistent accuracy increase without degradation
- **Fast Initial Learning:** Largest improvement occurs in first 1-2 batches
- **Stable Convergence:** System reaches target performance within 5-7 batches
- **No Overfitting:** Performance remains stable after convergence

DISCUSSION

Advantages of Hybrid Approach

The experimental results validate the effectiveness of the hybrid approach. The combination of rule-based and CRF strategies provides complementary strengths, addressing the limitations identified in recent work on document AI (Cui et al., 2021) and active learning (Settles, 2012):

- **Rule-based Strength:** Excels at structured fields with consistent formats (IDs, dates, phone numbers), achieving near-perfect accuracy with minimal computational cost.
- **CRF Strength:** Handles context-dependent fields with high variation (addresses, names, descriptions), learning patterns from limited feedback.
- **Adaptive Weighting:** Dynamic strategy selection based on confidence scores optimizes performance for each field type.
- **Fallback Mechanism:** Graceful degradation when primary strategy fails, maintaining high recall (99.97%).

The 35.74% relative improvement (25.96 percentage points absolute) validates our hypothesis that combining strategies yields superior performance. This gain stems from complementary strengths: rule-based excels at structured fields (IDs, dates) with minimal computation, while CRF handles context-dependent variations (names, addresses) through statistical learning. The intelligent selector optimizes per-field strategy choice based on confidence scores.

Data Efficiency as Core Contribution

The system's data efficiency represents a key contribution, addressing scenarios where extensive labeled datasets are unavailable. The system achieves 98.61% accuracy with only 25-35 documents per template, demonstrating effective learning from minimal data:

- **Minimal Data Requirement:** 25-35 documents per template (5-7 batches) to reach >95% accuracy, representing 10-40× less data than typical deep learning approaches
- **Real-World Alignment:** This scale matches realistic scenarios where small organizations have 20-50 documents per template type, not thousands
- **Few-Shot Learning Validation:** Demonstrates effective learning from minimal examples, following recent trends in data-scarce scenarios (Gebauer et al., 2023; Popovic & Färber, 2022)
- **Rapid Deployment:** Enables production deployment within days, not months of data collection and annotation

Positioning in the Landscape: Large-scale pre-trained models such as LayoutLMv3 (Huang et al., 2022) deliver excellent accuracy when ample labeled data and GPU resources are available. Our contribution complements those advances by demonstrating that a hybrid HITL system can reach 98.61% accuracy with only 25-35 documents per template and 7% correction effort in a resource-constrained, synthetic evaluation. Rather than replacing data-intensive pipelines, this work highlights a practical pathway for teams that need rapid deployment, limited annotation overhead, and CPU-only operation, leaving large-model integrations as future work once richer datasets become available.

HITL Learning Effectiveness

The Human-in-the-Loop learning mechanism proved highly effective:

- **Minimal User Burden:** Only 7% correction rate (196/2,800 fields) demonstrates practical feasibility for production deployment.
- **High Learning Efficiency:** 1.88 corrections per percentage point improvement is more efficient than traditional supervised learning requiring extensive labeling.
- **Fast Convergence:** System reaches >95% accuracy within 5-7 batches (25-35 documents), enabling rapid deployment.
- **Incremental Learning:** CRF model updates without full retraining, reducing computational cost and enabling real-time adaptation.

*name of corresponding author

The system demonstrates efficient learning characteristics: achieving 98.61% accuracy with only 7% field correction rate, yielding a learning efficiency score of 14.09 (calculated as final accuracy divided by correction rate: 98.61/7).

Generalization Capability

The system demonstrates strong generalization across document types:

- **Consistent Performance:** Standard deviation of only 2.19% across templates indicates stable performance.
- **Template Adaptability:** Successfully handles diverse structures from simple forms to complex mixed-format documents.
- **Scalability:** Architecture supports addition of new templates without redesign.

Template-specific analysis reveals:

- **Form & Table (100%):** Perfect accuracy achieved for structured templates
- **Letter (94.83%):** Most challenging due to high structural variation, but still strong performance
- **Mixed (99.59%):** Largest relative improvement (+43.25%), demonstrating hybrid approach effectiveness

Error Pattern Analysis: Detailed analysis of the most challenging template (Letter, 94.83% accuracy, 81 total errors) reveals systematic error patterns:

- **Partial/Truncated Extraction (64.2%):** The most common error type involves incomplete field extraction, particularly for address fields. Example: extracting "Jl" instead of "Jl. Pacuan Kuda No. 2 RT.065 RW.047". This occurs when extraction stops prematurely at punctuation or line breaks. HITL feedback effectively addresses this through pattern learning that captures complete field boundaries.
- **Context-Dependent Errors (27.2%):** Fields requiring contextual understanding pose challenges. Examples include extracting location names instead of dates (e.g., "Garut" vs "26-11-2024") or label text instead of field values (e.g., "Telp" vs "+62-064-092-8032"). CRF's contextual features help distinguish field content from surrounding text.
- **Prefix Text Contamination (3.7%):** Template-specific prefix text occasionally included in extraction. Example: "ini dibuat untuk keperluan X" instead of just "X". Pattern learning identifies and removes these prefixes through user corrections.
- **Missing Extraction (3.7%):** Some fields fail to extract any value, typically for fields with high positional variation or weak contextual cues. The adaptive system learns field locations through feedback.

This pattern validates the hybrid approach: rule-based handles structured fields effectively (near-perfect accuracy), while CRF's statistical learning is essential for context-dependent content. HITL feedback directly addresses the most common error types through incremental pattern learning and CRF adaptation. The error distribution shows that most issues (64.2%) stem from boundary detection rather than fundamental extraction failures, suggesting room for improvement through enhanced boundary features.

Threats to Validity

Several limitations should be considered when interpreting these results:

- **Template Diversity:** The evaluation covers 4 template types (Form, Table, Letter, Mixed). While this is sufficient to demonstrate cross-template generalization and validates the system's adaptability across diverse document structures, broader validation across more template types (e.g., scientific papers, legal documents, medical records) would strengthen generalizability claims. However, the few-shot learning capability and consistent performance across the evaluated templates suggest good generalization potential.
- **Synthetic Data:** The evaluation uses synthetically generated documents rather than real-world data from actual organizations. While this ensures reproducibility, perfect ground truth, and privacy compliance, validation with real documents would strengthen practical applicability claims. However, the synthetic generation simulates realistic document patterns, and the system's design (feature engineering, learning mechanisms) is not specific to synthetic data characteristics. The controlled experimental setup enables precise evaluation of core capabilities.
- **Dataset Scale:** The dataset size (140 documents, 2,800 fields) is intentionally designed to validate data efficiency rather than large-scale performance. This scale matches realistic deployment scenarios for small-to-medium organizations where large labeled datasets are unavailable. The key research question—"Can the system learn effectively from minimal data?"—is successfully demonstrated. While larger-scale evaluation would provide additional confidence, the current scale reflects the target use case: resource-constrained environments with limited labeled data.
- **Language Coverage:** Evaluation is limited to Indonesian documents. Cross-lingual validation would strengthen generalizability claims, though the feature engineering (layout, spatial, contextual) is largely language-independent.

*name of corresponding author

- **Baseline Comparison:** The baseline is rule-based only, which is appropriate for the target scenario (resource-constrained, data-scarce) but does not include GPU-based deep learning models. This choice reflects the practical constraint that GPU infrastructure is unavailable or cost-prohibitive in the target deployment environment. Comparison with LayoutLMv3/LLM is provided in Related Work as context rather than direct experimental comparison.
- **Template-Based Scope:** The system is designed for template-based documents with semi-structured layouts. Fully unstructured documents (e.g., academic papers, novels) are outside the current scope and would require different extraction approaches.

Limitations and Future Work

1) Current Limitations

- **Template Complexity:** Letter template (94.83% accuracy) shows room for improvement on highly variable structures with minimal consistent patterns.
- **Domain Specificity:** Focused on administrative documents; generalization to other domains (medical, legal) needs further research.
- **Cold Start:** Initial template requires some seed documents (5-10) for pattern learning; fully zero-shot extraction not supported.

2) Future Research Directions

Specific directions for extending this work include:

- **Hybrid Deep Learning Integration:** Combine the current lightweight CRF with transformer-based embeddings (e.g., DistilBERT) for improved contextual understanding while maintaining resource efficiency. This would leverage pre-trained language models for feature extraction while keeping the CRF layer for efficient incremental learning.
- **Transfer Learning Across Templates:** Implement meta-learning approaches to enable zero-shot or one-shot extraction on new templates by transferring knowledge from existing templates. This would address the cold-start limitation and enable immediate deployment on new document types.
- **Reinforcement Learning for HITL Optimization:** Apply reinforcement learning to optimize when and what to ask users, potentially reducing the 7% correction rate further by learning optimal interaction timing and sample selection strategies.
- **Multi-Modal Extension:** Incorporate visual features (document images) alongside text extraction, enabling handling of documents with complex visual layouts, tables with merged cells, and handwritten annotations that current text-based extraction cannot process.
- **Domain Adaptation:** Evaluate on specialized domains (medical records, legal contracts, scientific papers) to assess generalization beyond administrative documents and identify domain-specific challenges.
- **Online Learning in Production:** Implement continuous online learning during production deployment, enabling the system to adapt to evolving document formats and user preferences without explicit retraining cycles.

CONCLUSION

This research successfully developed an adaptive learning system based on Human-in-the-Loop (HITL) for PDF template data extraction that addresses critical gaps in current document processing approaches. The system integrates rule-based and CRF strategies in a unified hybrid framework with efficient incremental learning from user feedback.

Key Achievements

Experimental evaluation across diverse document types demonstrates the effectiveness and practical viability of the proposed hybrid approach. The integration of rule-based mechanisms with Conditional Random Field modeling produces substantial improvements in predictive accuracy while significantly reducing error rates, thereby confirming the complementary strengths of both strategies. The learning process remains highly efficient in practical settings, as the system requires only minimal user corrections to achieve meaningful performance gains, indicating strong feasibility for real-world deployment. Moreover, the model exhibits robust generalization capabilities, maintaining consistent performance across a wide range of document structures, from simple standardized forms to complex mixed-format documents, and achieving perfect accuracy in highly structured templates. The proposed architecture also demonstrates rapid adaptation, with fast convergence enabling efficient deployment on new document templates without the need for prolonged training cycles or large annotated datasets. Collectively, these results confirm that the system satisfies key production-level requirements by effectively

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

balancing accuracy, computational efficiency, and resource constraints, making it suitable for real-world operational environments.

Addressing Research Gaps

This research successfully addresses the three critical research gaps identified in the introduction through a unified and practical solution. First, it establishes a systematic hybrid human-in-the-loop framework that integrates rule-based methods for transparency with Conditional Random Fields for efficient machine learning within an adaptive learning context. The proposed architecture employs confidence-based dynamic weighting to harmonize both strategies, achieving an accuracy of 98.61% and demonstrating that high-performance adaptive systems can be realized without reliance on resource-intensive large-scale models. Second, the study designs and implements efficient feedback mechanisms that transform user corrections into actionable system knowledge via a dual-learning process, encompassing pattern learning for the rule-based component and incremental training for the CRF model. This approach yields high learning efficiency, requiring only 1.88 user corrections per percentage point of performance improvement and maintaining a low correction rate of 7%, thereby confirming the practicality of HITL for production deployment. Third, the research demonstrates real-time adaptive learning capabilities without extensive retraining by leveraging incremental CRF updates and rule-based pattern learning. The system consistently achieves production-ready accuracy levels exceeding 95% within just five to seven learning batches, corresponding to 25–35 documents, enabling rapid deployment and continuous performance improvement during real-world operation.

Research Contributions

Building on the identified research gaps, this study makes three substantive contributions to the field. From a methodological perspective, it introduces a novel hybrid architecture that systematically integrates rule-based transparency with Conditional Random Field statistical learning within a human-in-the-loop framework, thereby addressing the absence of coherent and structured hybrid approaches in existing literature. The proposed confidence-based adaptive weighting mechanism enables intelligent per-field strategy selection, allowing the system to dynamically leverage the strengths of both rule-based and statistical components. In terms of practical contribution, the research demonstrates a production-ready system capable of achieving competitive accuracy levels of up to 98.61% while requiring only minimal training data, limited user intervention, and moderate computational resources. By operating effectively with as few as 25–35 documents, maintaining a low user correction rate of 7%, and relying solely on CPU-based processing with modest memory requirements, the proposed approach effectively bridges the gap between rigid rule-based systems and resource-intensive large-scale models. Empirically, the study provides a comprehensive evaluation across 140 documents and four diverse template types, offering robust evidence that hybrid human-in-the-loop architectures can deliver high accuracy alongside practical deployability. These results validate both the generalization capability and real-world viability of the proposed approach.

Implications

The research has significant implications for both academia and industry: **For Researchers:** Demonstrates that efficient hybrid architectures can achieve competitive accuracy while maintaining resource efficiency and adaptive learning capability—offering a practical alternative to resource-intensive approaches. **For Practitioners:** Provides a viable solution for document extraction automation with clear deployment path, quantifiable business value, and manageable implementation requirements. **For HITL Systems:** Shows that human-in-the-loop can be more than validation—it can be the primary driver of efficient incremental learning when properly integrated into system architecture.

Broder Impact

This research advances the field of document extraction by demonstrating that high accuracy and practical deployability are not mutually exclusive. The hybrid HITL approach opens new possibilities for organizations and domains previously unable to adopt automated extraction due to resource or data constraints. By achieving production-ready performance with minimal requirements, this work democratizes access to intelligent document processing technology. The system's design philosophy—prioritizing data efficiency, resource frugality, and incremental learning—offers a template for developing AI systems that serve real-world needs beyond benchmark performance. This is particularly relevant as the field increasingly recognizes the importance of sustainable, accessible, and practical AI solutions.

Future Vision

Building on this foundation, future research should explore: **Enhanced Adaptability:** Developing zero-shot and few-shot capabilities to eliminate cold-start requirements entirely, enabling immediate deployment without

*name of corresponding author

seed documents. **Broader Applicability:** Extending the approach to diverse domains beyond administrative documents, including scientific, medical, and legal document processing. **Ecosystem Development:** Building tools, standards, and best practices to facilitate wider adoption, including standardized interfaces and integration patterns for practitioners.

This research demonstrates that intelligent hybrid architectures can bridge the gap between traditional and modern approaches to document extraction. By combining the strengths of rule-based transparency with CRF adaptability in a Human-in-the-Loop framework, the system achieves the practical balance needed for real-world deployment: high accuracy with minimal data, low resources, and manageable user effort. This work contributes not only a functional system but also a methodology for developing AI solutions that prioritize accessibility and practical impact alongside technical performance.

ACKNOWLEDGEMENT

The authors would like to thank Universitas Nasional for providing the research facilities and support. We also thank the anonymous reviewers for their valuable feedback that helped improve this manuscript.

DATA AVAILABILITY

The experimental data and code implementation used in this study are available from the corresponding author upon reasonable request, subject to institutional data protection policies.

REFERENCES

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445717

Chiticariu, L., Li, Y., & Reiss, F. R. (2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 827–832. https://doi.org/10.18653/v1/d13-1079

Cui, L., Xu, Y., Lv, T., & Wei, F. (2021). *Document AI: Benchmarks, Models and Applications* (No. arXiv:2111.08609). arXiv. https://doi.org/10.48550/arXiv.2111.08609

Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/TPAMI.2021.3057446

Dudley, J. J., & Kristensson, P. O. (2018). A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems*, *8*(2), 1–37. https://doi.org/10.1145/3185517

Garncarek, Ł., Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Turski, M., & Graliński, F. (2021). LAMBERT: Layout-Aware Language Modeling for Information Extraction. In J. Lladós, D. Lopresti, & S. Uchida (Eds.), *Document Analysis and Recognition – ICDAR 2021* (Vol. 12821, pp. 532–547). Springer International Publishing. https://doi.org/10.1007/978-3-030-86549-8_34

Gebauer, M., Maschhur, F., Leschke, N., Grünewald, E., & Pallas, F. (2023). A 'Human-in-the-Loop' approach for Information Extraction from Privacy Policies under Data Scarcity. *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 76–83. https://doi.org/10.1109/EuroSPW59978.2023.00014

Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., & Park, S. (2022). BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 10767–10775. https://doi.org/10.1609/aaai.v36i10.21322

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 4083–4091. https://doi.org/10.1145/3503161.3548112

International Organization for Standardization. (2008). *Document management—Portable document format—Part 1: PDF 1.7* (No. ISO 32000-1:2008). ISO. https://www.iso.org/standard/51502.html

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270. https://doi.org/10.18653/v1/N16-1030

Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., & Ding, E. (2021). StrucTexT: Structured Text Understanding with Multi-Modal Transformers. *Proceedings of the 29th ACM International Conference on Multimedia*, 1912–1920. https://doi.org/10.1145/3474085.3475345

*name of corresponding author

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Munro, R. (with Safari, an O'Reilly Media Company). (2021). *Human-in-the-Loop Machine Learning* (1st edition). Manning Publications.

Palm, R. B., Winther, O., & Laws, F. (2017). CloudScan—A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 406–413. https://doi.org/10.1109/icdar.2017.74

Popovic, N., & Färber, M. (2022). Few-Shot Document-Level Relation Extraction. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5733–5746. https://doi.org/10.18653/v1/2022.naacl-main.421

Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., & Pałka, G. (2021). *Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer* (No. arXiv:2102.09550). arXiv. https://doi.org/10.48550/arXiv.2102.09550

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2022). A Survey of Deep Active Learning. *ACM Computing Surveys*, *54*(9), 1–40. https://doi.org/10.1145/3472291

Schleith, J., Hoffmann, H., Norkute, M., & Cechmanek, B. (2022). *Human in the loop information extraction increases efficiency and trust*. https://doi.org/10.18420/MUC2022-MCI-WS12-249

Schroeder, N. L., Jaldi, C. D., & Zhang, S. (2025). *Large Language Models with Human-In-The-Loop Validation for Systematic Review Data Extraction* (No. arXiv:2501.11840). arXiv. https://doi.org/10.48550/arXiv.2501.11840

Settles, B. (2012). *Active Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-031-01560-1

Wu, T., Terry, M., & Cai, C. J. (2022). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3491102.3517582

*name of corresponding author