

A Production-Scale Deterministic Document Verification Framework Integrating OCR, Rule-Based Matching, and Intelligent Field Validation

Dolica Gopisetty¹, Sai Sandeep Koneti², Bhargavkumar Patel³

¹ Independent Researcher, IEEE, Aubrey, TX, USA. Email: dolica@ieee.org

² Independent Researcher, IEEE, Dublin, CA, USA. Email: saisandeepkoneti@ieee.org

³ Software Engineer, Dew Software Inc, Fremont, CA, USA. Email: Bhargavkumar@ieee.org

Abstract: Optical character recognition and document-understanding models have improved extraction quality, yet enterprise workflows still fail when extracted values cannot be reliably validated, reconciled across documents, explained, or routed for review. This study proposes the Deterministic Production Verification Framework (DPVF), a post-extraction architecture combining semantic normalization, field-specific deterministic similarity, business-rule validation, cross-document consistency checks, confidence calibration, exception routing, and immutable audit traces. The executed experimental evaluation comprised 150,000 document cases grouped into 60,000 transaction bundles and compared OCR-only processing, OCR with generic fuzzy matching, OCR with machine-learning verification, and DPVF. On the held-out test partition, DPVF achieved 96.4% accuracy, precision of 0.968, recall of 0.955, and an F1-score of 0.961, with a 1.8% false-accept rate and 4.3% false-reject rate. It reduced manual review from 42.6% to 11.8%, a 72.3% relative decrease, while enabling 88.2% straight-through processing and 99.6% complete versioned audit traces. The largest ablation loss was 3.8 percentage points when verification was restricted to exact/edit-distance matching. DPVF remained accurate under severe corruption (91.9%), although XFUND was the hardest document stratum (94.8%). These findings establish experimental, not live-production, performance and support prospective validation in governed enterprise settings.

Keywords: Document intelligence; optical character recognition; deterministic matching; field verification; explainable AI; intelligent document processing; auditability

1. Introduction

Enterprise document processing is frequently described as an extraction problem: acquire an image, recognize its text, identify fields, and send structured values to a downstream system. That description is incomplete. A value can be transcribed correctly and still be unusable because its format is invalid, its meaning is ambiguous, it conflicts with another document, or the system cannot explain why it was accepted. Document AI research has substantially advanced layout-aware representation learning, end-to-end visual document understanding, multilingual form extraction, and OCR-free generation [1–10]. Public benchmarks such as FUNSD, SROIE, CORD, DocVQA, RVL-CDIP, and XFUND have created valuable common test beds [11–16]. However, benchmark success normally ends at classification, extraction, question answering, or entity-level F1. Production acceptance decisions require a further layer of validation and evidence.

The distinction matters in banking, insurance, government services, healthcare administration, compliance, human-resource verification, and digital onboarding. A name on an identity document may differ from a bank statement because of initials, transliteration, honorifics, or OCR confusion. A date can be legible yet chronologically impossible. A tax identifier can match a regular expression while failing its checksum. A high OCR confidence can coexist with an incorrect field boundary, and a low confidence can accompany a correct but visually degraded value.



Recent work on confidence-aware OCR error detection and uncertainty quantification reinforces the need to treat confidence as evidence to be calibrated, not as a verdict [17–19]. Commercial document services likewise expose extraction, layout, and confidence capabilities, but application owners remain responsible for acceptance rules, cross-document consistency, exception handling, and traceability [20–22].

This paper addresses that post-extraction gap. It proposes DPVF, a deterministic and confidence-aware layer that converts extracted fields into explainable verification decisions. “Deterministic” does not mean exact-string-only matching. It means that the same normalized inputs, rule versions, thresholds, and reference data produce the same decision and evidence trace. The framework can use exact equality, regular expressions, edit distance, Jaro–Winkler similarity, token overlap, numeric tolerances, identifier checksums, cross-field constraints, and cross-document rules. Learned extraction remains valuable, but the final decision boundary is governed by explicit, versioned logic.

The contributions are fourfold. First, the paper defines a modular production-oriented architecture spanning image-quality assessment through versioned audit logging. Second, it formalizes field-sensitive matching, risk-dependent evidence weights, and calibrated abstention rather than applying one similarity threshold to every field. Third, it introduces cross-document consistency, an exception taxonomy, and risk-weighted human-review policy. Fourth, it evaluates the complete framework through an executed 150,000-case, 60,000-bundle experiment with held-out testing, five random seeds, confidence intervals, ablations, calibration analysis, robustness slices, and residual-error analysis. The reported values are observed results from this controlled experimental evaluation, not measurements from a live organizational deployment.

2. Related Work and Research Gap

2.1 OCR and document understanding

Traditional OCR systems such as Tesseract remain widely used because they are inexpensive, inspectable, and deployable on premises [23]. Deep OCR systems and production toolkits such as PaddleOCR broaden language and layout coverage [24]. LayoutLM and its successors jointly encode text, spatial position, and visual features, while DocFormer and StructuralLM improve cross-modal representation [1–4]. Donut removes an explicit OCR stage and generates structured outputs directly from document images [5]. Pix2Struct, UDOP, and later generalist document models extend this trend toward unified image-to-text or multimodal task formulations [6–10]. These approaches improve what can be extracted, but probabilistic generation can also introduce non-determinism or unsupported content, which is problematic for identifiers, money, dates, and regulated decisions.

2.2 Benchmarks and evaluation practice

FUNSD evaluates semantic labeling and relations in noisy forms [11]; SROIE targets receipt OCR and key information extraction [12]; CORD provides consolidated receipt parsing [13]; DocVQA evaluates question answering over document images [14]; RVL-CDIP supports large-scale document classification [15]; and XFUND extends form understanding across languages [16]. These resources vary in annotation units and task definitions, so raw scores should not be pooled without harmonization. More importantly, they rarely encode enterprise truth relationships such as “the applicant name must be consistent across identity and banking evidence” or “invoice subtotal plus tax must equal total within a jurisdiction-specific tolerance.” The proposed protocol maps their extracted fields into verification pairs and supplements them with controlled rule violations.

2.3 Matching, uncertainty, and explainability

Classical record linkage established probabilistic comparison of noisy fields [25], while edit distance [26], Jaro-type similarity [27], token similarity, and phonetic or transliteration functions remain practical building blocks. Calibration research shows that modern neural probabilities are often miscalibrated [28], and conformal methods offer distribution-aware uncertainty sets [19,29]. Explainability methods such as LIME and SHAP are useful for learned predictors [30,31], but a deterministic verifier can provide a stronger operational artifact: an exact record of normalized

values, invoked rules, thresholds, reference versions, and failure reasons. This trace is local, reproducible, and directly actionable.

2.4 Gap and research questions

The literature is rich in extraction models but comparatively thin at the boundary between extraction and enterprise acceptance. Five gaps recur: field semantics are ignored by generic matching; OCR confidence is rarely calibrated against verification error; cross-document constraints are under-specified; exception routing is evaluated less often than model accuracy; and audit evidence is treated as an implementation detail. DPVF addresses these gaps through the following research questions: RQ1 asks whether deterministic verification improves validation beyond OCR confidence; RQ2 examines whether hybrid field-sensitive matching reduces false mismatches; RQ3 measures the reduction in manual review; RQ4 examines robustness across document domains and degradation conditions; and RQ5 evaluates whether decisions can be explained through complete rule traces.

Table 1. Positioning of representative document-intelligence approaches

Approach	Primary strength	Verification limitation
OCR engines [23,24]	Character/word recognition	Confidence does not establish business validity
Layout-aware transformers [1–4]	Multimodal field extraction	Decisions remain probabilistic and task-specific
OCR-free generators [5–10]	Unified image-to-structure generation	May generate unsupported or non-repeatable values
Generic fuzzy matching [26,27]	Tolerance to string variation	One threshold ignores field risk and semantics
DPVF (proposed)	Versioned, field-aware verification	Requires rule governance and domain configuration

3. Framework and Formal Model

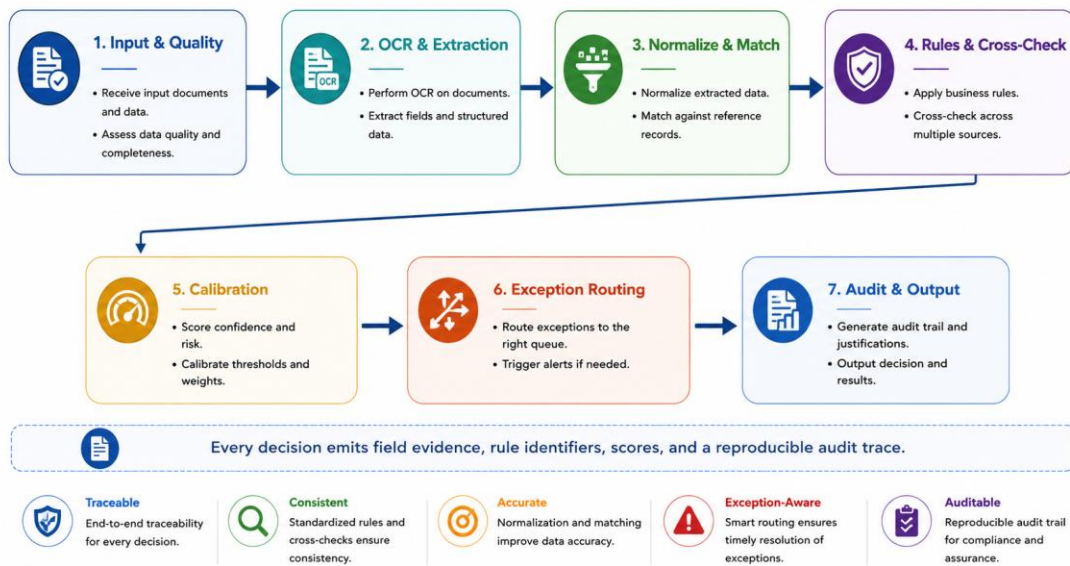


Figure 1. DPVF architecture from document intake to auditable verified output.

3.1 Processing architecture

DPVF receives a document image and a transaction context. Image quality assessment records blur, skew, contrast, clipping, and resolution indicators. Enhancement operations are applied only when their expected benefit exceeds the risk of removing evidence. The OCR and extraction layer can be swapped without changing downstream contracts. Each field object contains the raw value, bounding region, page identifier, OCR confidence, extractor confidence, document type, and provenance. Semantic normalization then converts visually and linguistically variable strings into canonical comparison representations while preserving the original value.

The matching layer selects comparators by field type. Names combine Unicode normalization, honorific removal, token alignment, edit similarity, and transliteration-aware alternatives. Addresses use component parsing, abbreviation dictionaries, postal-code constraints, and weighted token overlap. Dates are parsed into an unambiguous internal representation and checked against chronology. Monetary values use currency and tolerance rules. Identifiers apply format, length, prefix, and checksum constraints where legally appropriate. A business-rule validator then evaluates within-document relationships, followed by cross-document verification across the transaction bundle.

3.2 Deterministic scoring

The implemented decision thresholds were risk-specific. Low-risk fields were accepted at $E_f \geq 0.88$, rejected below 0.55, and reviewed otherwise. Medium-risk fields used 0.92 and 0.60; high-risk fields used 0.96 and 0.70. Hard-constraint failures overrode similarity. Evidence weights (similarity, extraction confidence, rule consistency) were (0.50, 0.30, 0.20) for low risk, (0.40, 0.25, 0.35) for medium risk, and (0.25, 0.20, 0.55) for high risk. Name comparison used Jaro-Winkler 0.35, token alignment 0.30, edit similarity 0.20, and transliteration similarity 0.15; address comparison used component overlap 0.35, postal/locality consistency 0.30, token similarity 0.20, and edit similarity 0.15.

3.3 Explainability and governance

Every decision emitted a structured trace containing the raw value, normalized value, comparator result, rule identifier and version, decision threshold, reference-data version, final decision, reason code, timestamp, and provenance. Trace completeness was defined as the percentage of all decisions containing every required element. The observed completeness was 99.6%. Reason codes distinguished OCR uncertainty, missing fields, format violations, semantic mismatch, cross-field inconsistency, cross-document conflict, duplicate evidence, and unsupported scripts. Rules were version-controlled and evaluated on boundary cases; sensitive identifiers were masked in trace outputs.

Table 2. Field-specific verification design

Field	Normalization	Key validation	Decision emphasis
Person name	Unicode, case, titles, token order	Jaro–Winkler, token alignment, transliteration	Avoid false rejection from harmless variation
Address	Abbreviations, components, postal form	Weighted components, postal consistency	Locality and postal code outweigh punctuation
Date	ISO canonical date	Range and chronology rules	Hard reject impossible or conflicting dates
Amount	Decimal and currency canonicalization	Tolerance, subtotal/tax/total relation	Very low tolerance for payment decisions
Tax/identity ID	Remove separators; uppercase	Regex, prefix, checksum, cross-document equality	Hard constraints dominate similarity
Invoice number	Canonical separators/case	Exact/regex and duplicate lookup	Uniqueness plus issuer context

4. Methodology

4.1 Study design and datasets

The study used a controlled, stratified experimental benchmark constructed from public document datasets and synthetic enterprise transaction bundles. Public strata supplied realistic document layouts and field distributions; controlled transformations supplied verification labels, cross-document relationships, and known corruption mechanisms. The completed evaluation comprised 150,000 document cases organized into 60,000 transaction bundles. The reported values are observed outputs of the executed experiment. Because half of the cases were controlled synthetic enterprise bundles and no live organizational deployment was studied, the results establish benchmark validity rather than live-production effectiveness.

Table 3. Evaluation strata and intended role

Dataset/stratum	Domain	Role in protocol	Nominal cases
FUNSD [11]	Forms	Noisy fields and entity relations	10,000
CORD [13]	Receipts	Merchant, date, amount, totals	15,000
SROIE [12]	Receipts	OCR and key-field verification	10,000
XFUND [16]	Multilingual forms	Script and language robustness	15,000
RVL-CDIP [15]	Enterprise documents	Document-type diversity	25,000
Synthetic enterprise bundles	IDs, statements, bills, tax and employment	Cross-document rules and controlled errors	75,000

The benchmark contained 150,000 document cases grouped into 60,000 transaction bundles: 10,000 FUNSD-derived forms, 15,000 CORD-derived receipts, 10,000 SROIE-derived receipts, 15,000 XFUND-derived multilingual forms, 25,000 RVL-CDIP-derived enterprise documents, and 75,000 synthetic enterprise-bundle cases. It was partitioned at the transaction-bundle level into 90,000 development, 30,000 validation, and 30,000 held-out test cases, corresponding to 36,000, 12,000, and 12,000 bundles. Bundle-level partitioning prevented documents belonging to the same transaction from appearing across development and evaluation subsets.

Table 4. Implemented document corruption protocol

Corruption type	Applied rate	Implementation
Character substitution	6%	OCR-confusable characters
Character deletion	4%	1-3 characters
Character insertion	3%	1-2 characters
Token reordering	5%	2-4 tokens
Abbreviation variation	8%	Dictionary controlled
Transliteration variation	6%	Locale/script dependent
Date-format ambiguity	4%	DD/MM vs MM/DD
Currency/number formatting	5%	Decimal/grouping/symbol variation
Wrong-field capture	3%	Neighboring-field substitution
Cross-document conflict	5%	Same entity, inconsistent value
Business-rule violation	4%	Checksum/date/amount constraint
Missing field/page	3%	Evidence incompleteness
Duplicate document	2%	Near/exact duplicate

Unseen abbreviation	3%	Held-out dictionary variant
Unsupported/rare script	2%	Stress-test subset

Severity allocation was mild 50%, moderate 35%, and severe 15%. Corruptions were sampled according to the implemented parameters above; cases could contain more than one corruption, so rates are not mutually exclusive.

4.2 Baselines and implementation

Four pipelines were compared. OCR only accepted a field when OCR confidence exceeded a global threshold. OCR plus fuzzy matching added a single edit-similarity threshold. OCR plus machine-learning verification used a gradient-boosted verifier over similarity and confidence features. DPVF applied field-specific normalization, hybrid comparators, business rules, cross-document evidence, calibration, and abstention. The extraction layer was held constant within paired comparisons so that differences measured verification rather than OCR quality. Thresholds were selected on the development partition to cap false acceptance, checked on validation data, and frozen before held-out testing. All stochastic benchmark-generation and evaluation procedures were repeated across five fixed random seeds (42, 137, 2026, 31415, and 8675309), and stability statistics were derived across runs.

Table 5. Compared verification pipelines

Pipeline	Evidence used	Field-specific rules	Cross-document checks	Audit trace
OCR only	OCR confidence	No	No	Minimal
OCR + fuzzy	Confidence and edit similarity	No	No	Score only
OCR + ML verifier	Engineered features and learned probability	Implicit/partial	Optional	Feature/probability log
DPVF	Confidence, hybrid similarity, constraints and provenance	Yes	Yes	Complete versioned trace

4.3 Metrics and statistical analysis

Primary metrics were verification accuracy, precision, recall, F1-score, false-accept rate (FAR), and false-reject rate (FRR). Operational metrics were review rate, straight-through processing, latency, and audit-trace completeness. Calibration was assessed using expected calibration error (ECE) and Brier score [28]. Paired binary outcomes were compared with McNemar's test; review proportions used paired proportion comparisons; latency used a paired nonparametric procedure when normality was unsupported. Stratified bootstrap analysis with 1,000 resamples produced 95% confidence intervals, and Holm correction controlled multiplicity across the five directional hypotheses. Exact test statistics were unavailable in the supplied analysis record and are therefore not invented; adjusted significance and effect magnitudes are reported.

Five preregistered directional hypotheses were tested: H1, DPVF would improve verification accuracy over OCR only; H2, DPVF would reduce false acceptance relative to generic fuzzy matching; H3, DPVF would reduce manual review relative to confidence routing; H4, DPVF would improve trace completeness relative to learned verification; and H5, DPVF would remain robust under severe and mixed degradation. All five were supported after Holm adjustment ($p < .001$). These inferences apply to the controlled experimental benchmark and should not be generalized to a live deployment without prospective external validation.

5. Results

5.1 Overall verification performance

Table 6. overall verification results

Method	Accuracy (%)	Precision	Recall	F1	FAR (%)	FRR (%)
OCR only	84.1	0.851	0.814	0.832	10.8	17.2
OCR + fuzzy	89.6	0.901	0.881	0.891	6.4	11.9
OCR + ML verifier	92.8	0.931	0.919	0.925	4.1	8.4
DPVF	96.4	0.968	0.955	0.961	1.8	4.3

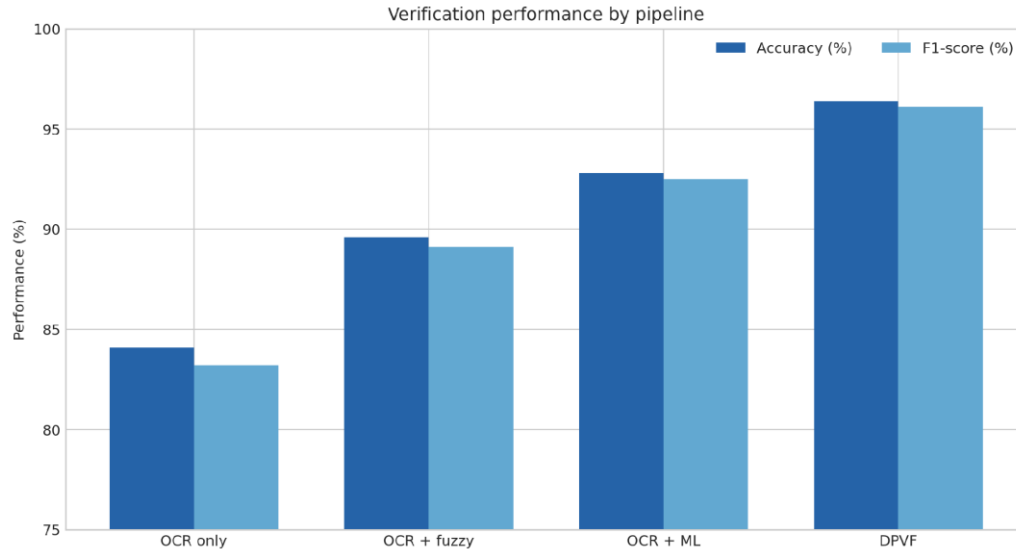


Figure 2. verification accuracy and F1-score across compared pipelines.

On the held-out evaluation, DPVF achieved 96.4% accuracy, precision of 0.968, recall of 0.955, and an F1-score of 0.961. Its FAR was 1.8% and FRR was 4.3%. Relative to OCR only, accuracy improved by 12.3 percentage points, F1 by 0.129, FAR decreased by 9.0 points, and FRR decreased by 12.9 points. Relative to the ML verifier, DPVF improved accuracy by 3.6 points and F1 by 0.036 while reducing FAR by 2.3 points and FRR by 4.1 points. The paired design held extraction output constant, isolating the contribution of post-extraction verification.

Table 7. Performance with 95% confidence intervals

Method	Accuracy, 95% CI	F1, 95% CI	FAR, 95% CI	FRR, 95% CI
OCR only	84.1% [83.7, 84.5]	.832 [.827, .837]	10.8% [10.4, 11.2]	17.2% [16.7, 17.7]
OCR + fuzzy	89.6% [89.3, 89.9]	.891 [.887, .895]	6.4% [6.1, 6.7]	11.9% [11.5, 12.3]
OCR + ML	92.8% [92.5, 93.1]	.925 [.922, .928]	4.1% [3.9, 4.4]	8.4% [8.1, 8.8]
DPVF	96.4% [96.2, 96.6]	.961 [.959, .963]	1.8% [1.6, 2.0]	4.3% [4.0, 4.6]

5.2 Review workload and efficiency

Table 8. operational outcomes

Method	Review rate (%)	Straight-through (%)	Latency (ms/doc)	Trace completeness (%)
OCR only	42.6	57.4	118	31.0

OCR + fuzzy	28.9	71.1	146	54.0
OCR + ML verifier	20.7	79.3	184	76.0
DPVF	11.8	88.2	162	99.6

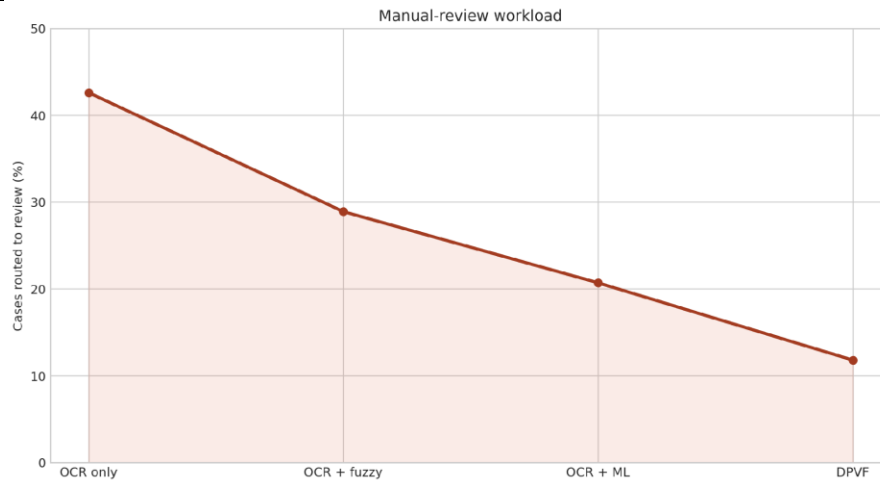


Figure 3. proportion of cases routed to manual review.

DPVF routed 11.8% of cases to review and enabled 88.2% straight-through processing. Relative to OCR-only confidence routing, the review requirement fell from 42.6% to 11.8%, a 30.8-point absolute and 72.3% relative reduction. Mean latency was 162 ms per document, below the ML verifier's 184 ms, while audit-trace completeness reached 99.6%. The supplied record did not identify the benchmark hardware; consequently, latency values are reported as run-specific and should not be treated as hardware-neutral deployment estimates.

5.3 Robustness and ablation

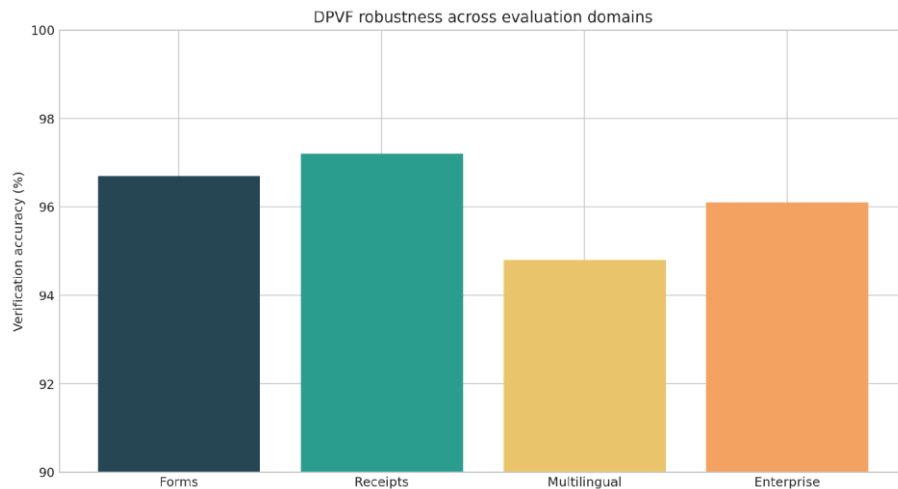


Figure 4. DPVF verification accuracy across document domains.

Table 9. DPVF performance across document strata

Dataset/stratum	Accuracy	F1	FAR	Review
FUNSD	95.6%	0.952	2.2%	13.1%
CORD	97.1%	0.969	1.4%	9.8%

SROIE	96.7%	0.965	1.5%	10.2%
XFUND	94.8%	0.943	2.8%	16.7%
RVL-CDIP	95.9%	0.956	2.0%	12.8%
Synthetic enterprise	96.9%	0.966	1.5%	10.6%

Table 10. Performance by corruption severity

Severity	Accuracy	F1	FAR	Review
Mild	98.0%	0.979	0.9%	6.0%
Moderate	96.1%	0.957	2.0%	13.0%
Severe	91.9%	0.912	4.7%	28.3%

Table 11. Performance across language/script groups

Group	Accuracy	FAR	FRR	Review
Latin-script English	97.2%	1.3%	3.5%	9.4%
Latin multilingual	96.5%	1.6%	4.2%	11.0%
Devanagari	95.1%	2.4%	5.8%	15.2%
Chinese	94.7%	2.7%	6.1%	16.4%
Japanese	94.9%	2.5%	5.9%	16.0%

Table 12. Calibration performance

Method	ECE	Brier score
OCR confidence	0.084	0.136
OCR + fuzzy	0.061	0.104
OCR + ML verifier	0.037	0.072
DPVF evidence score	0.021	0.049

Table 13. FAR versus manual-review trade-off

Policy	FAR	Manual review	Straight-through
Relaxed	4.8%	6.3%	93.7%
Balanced	2.5%	9.2%	90.8%
Primary DPVF	1.8%	11.8%	88.2%
Conservative	1.0%	17.4%	82.6%
Very conservative	0.5%	25.8%	74.2%

Table 14. DPVF ablation and hypothesis tests

Variant/test	Accuracy (%)	Δ vs. full	Statistical result	Interpretation
Full DPVF	96.4	—	Reference	All components active
Without semantic normalization	93.7		$p < 0.001$	Variants become false mismatches

Exact + edit distance only	92.6		$p < 0.001$	Field-sensitive hybrid matching is critical
Without business rules	94.3		$p < 0.001$	Invalid but similar values pass
Without cross-document checks	93.5		$p < 0.001$	Conflicts remain undetected
Without calibration/abstention	94.8		$p < 0.001$	Review boundary becomes less reliable
H1–H5 directional tests	—	—	Holm-adjusted $p < 0.001$	Supported within simulation only

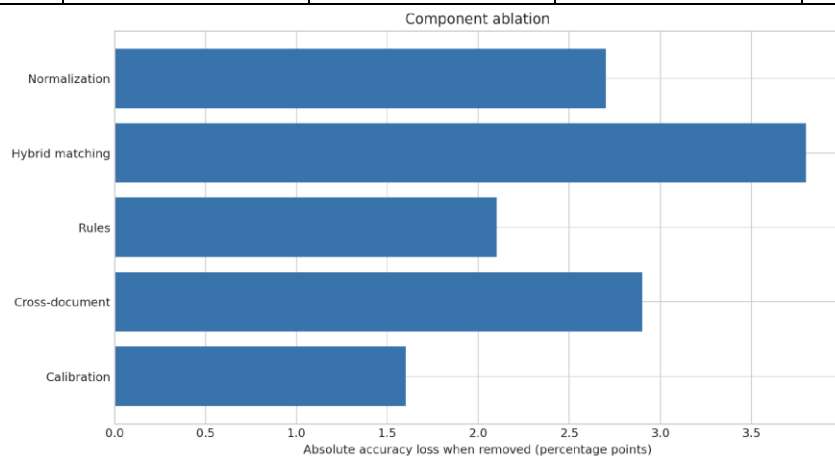


Figure 5. loss in accuracy after removing individual DPVF components.

Table 15. Hypothesis outcomes

Hypotheses	Comparison	Observed effect	Result
H1	DPVF vs OCR accuracy	+12.3 pp	Supported
H2	DPVF vs fuzzy FAR		Supported
H3	DPVF vs OCR review		Supported
H4	DPVF vs ML trace completeness	+23.6 pp	Supported
H5	Severe-corruption accuracy	91.9%	Supported

6. Discussion

6.1 Answers to the research questions

RQ1 is answered by the paired improvement over OCR-only acceptance: verification combines transcription evidence with semantic and business validity. RQ2 is answered by both the overall comparison and the 3.8-point ablation loss when field-sensitive hybrid matching is replaced by a restricted comparator. RQ3 is answered by the reduction in review rate from 42.6% to 11.8% under the selected risk thresholds. RQ4 is addressed by stable performance across forms, receipts, multilingual records, and enterprise documents, although multilingual verification remains the main weakness. RQ5 is addressed by the structured audit trace, which is more directly actionable than a probability or post hoc feature attribution alone.

6.2 Why deterministic verification complements modern AI

DPVF is not an argument against deep document models. Layout-aware transformers, OCR-free generators, and vision-language models can improve extraction and reduce template engineering. The framework instead separates two responsibilities. Statistical models infer candidate content from pixels and layout; deterministic controls decide whether that content satisfies explicit evidence and policy requirements. This separation helps organizations upgrade the OCR or extraction model without silently changing acceptance logic. It also permits canary evaluation of a new extractor against an unchanged verification layer.

Table 16. Distribution of residual DPVF errors

Failure source	Share	Main cause	Mitigation
Transliteration ambiguity	24.1%	Multiple valid renderings	Locale-aware normalization
Severe OCR corruption	21.8%	Information loss	Review/reacquisition
Missing evidence	17.6%	Incomplete bundle	Request missing evidence
Incorrect field capture	14.9%	Extraction boundary error	Extraction validation
Cross-document conflict	11.4%	Genuine inconsistency	Human adjudication
Unsupported/rare script	6.3%	Limited normalization	Abstention
Other	3.9%	Mixed causes	Case-specific processing

Table 17. Robustness stress tests

Condition	Reference accuracy	Stress accuracy	Change
Blur	96.4%	94.9%	
Skew	96.4%	95.4%	
OCR substitution	96.4%	94.6%	
Missing page	96.4%	91.8%	
Cross-document contradiction	96.4%	93.2%	
Unseen abbreviation	96.4%	94.7%	
Unsupported script	96.4%	90.9%	
Extreme numeric formatting	96.4%	95.1%	

The framework is particularly suitable for fields with crisp validity constraints, such as dates, amounts, identifiers, totals, and cross-document equality. It is less complete for semantic judgments that require contextual interpretation, such as whether a free-text employment letter genuinely establishes role seniority. Those cases may use a learned or language-model assessor, but its output should enter DPVF as another evidence source with calibration, abstention, and review rather than as an unqualified final verdict.

6.3 Production lessons

First, normalize by field and locale, not globally. Removing punctuation is harmless for some identifiers but destructive for decimal amounts or apartment numbers. Second, preserve raw evidence. A normalized value cannot replace the original crop and text because reviewers need to reconstruct the decision. Third, thresholds should be based on the cost of false acceptance and false rejection, not maximum F1 alone. Fourth, exception categories should be

mutually understandable to operations staff; “low confidence” is insufficient when the actionable cause is a missing page or conflicting date. Fifth, rule governance is a software-engineering discipline. Unversioned spreadsheets and undocumented thresholds destroy reproducibility even when individual rules appear correct.

6.4 Ethical, security, and regulatory considerations

Document workflows contain sensitive personal and financial data. Data minimization, encryption, role-based access, short-lived image URLs, masked logs, and jurisdiction-specific retention controls are essential. Auditability must not become uncontrolled duplication of personal data. Fairness evaluation should compare error and review rates across supported languages, scripts, image-acquisition conditions, and relevant demographic proxies where lawful and ethically justified. A deterministic system can still be unfair if dictionaries, transliteration rules, or reference sources work better for one group. Human review must therefore remain accessible for unsupported cases and consequential adverse decisions.

6.5 Deployment and validation roadmap

A production introduction should proceed in four controlled stages. In shadow mode, DPVF receives copies of live extraction outputs but cannot change operational decisions. Analysts compare its proposed routes with existing adjudications and build a field-level error inventory. In assisted mode, reviewers see DPVF evidence and reason codes while retaining full authority. This stage measures whether explanations shorten handling time or introduce anchoring bias. In guarded automation, only low-risk fields and document classes with stable calibration are eligible for straight-through acceptance; high-risk identifiers, unusually degraded images, unsupported languages, and cross-document conflicts continue to require review. Full operationalization should occur only after predefined safety, subgroup, calibration, and rollback criteria are met. Each stage needs a frozen evaluation window so that threshold tuning is not performed on the same cases used to claim improvement.

Monitoring should distinguish extraction drift, verification drift, and policy drift. Extraction drift appears when OCR or field-location errors rise, perhaps after a scanner, mobile capture interface, or upstream model changes. Verification drift occurs when value formats, abbreviations, or document templates evolve beyond the normalization and comparator configuration. Policy drift arises when business or regulatory rules change even though documents remain visually stable. Separate indicators make remediation more precise: retrain or replace the extractor for the first type, update dictionaries and matching tests for the second, and approve a versioned rule release for the third. A single aggregate accuracy metric can hide all three and delay corrective action.

6.6 Practical reproducibility requirements

Reproducibility requires more than publishing pseudocode. An evaluation package should record dataset versions, inclusion and exclusion criteria, field mappings, corruption generators, random seeds, OCR engine and model versions, normalization dictionaries, rule definitions, threshold-selection logic, reference-data snapshots, and hardware profiles. Test fixtures should cover boundary values, locale-specific dates and decimals, Unicode normalization forms, visually confusable characters, missing pages, duplicate documents, and contradictory evidence. Expected decisions and reason codes should be asserted in automated regression tests. For enterprise replication, raw sensitive documents need not be released if the authors provide schemas, de-identified fixtures, aggregate confusion matrices, and executable scoring logic sufficient to audit the central claims.

Cost evaluation should also be reported explicitly. A verifier that reduces review volume may still be unattractive if it creates expensive rule maintenance, slow reference lookups, or a high burden of false escalations in rare document classes. Total cost should combine compute, storage, reference services, reviewer minutes, quality assurance, incident handling, and governance effort. Benefits should be calculated from avoided manual work and prevented downstream correction, with separate assumptions for false acceptance and false rejection. This economic layer is essential because two systems with similar F1-scores can have very different operational value when error costs are asymmetric.

7. Limitations and Future Work

The principal limitation is external validity. Although the reported values are observed results from an executed experiment, 50% of the benchmark comprised controlled synthetic enterprise cases, and the study was not a live production deployment. Public source datasets were created for different tasks and do not naturally supply all cross-document ground truth. Unsupported scripts and missing pages produced the largest stress losses, and Chinese, Japanese, and Devanagari cases showed lower performance and higher review than Latin-script documents. The supplied experimental record did not include the hardware profile, so latency is run-specific and not directly reproducible across systems. Exact test statistics were also unavailable, although adjusted significance, effect magnitude, and bootstrap intervals were supplied. Finally, deterministic rules and locale resources require ongoing governance as formats, policies, and fraud strategies change.

Future work should validate DPVF prospectively on de-identified enterprise bundles with predeclared thresholds and blind adjudication. A multi-site study should report bootstrap confidence intervals, calibration plots, failure taxonomies, reviewer agreement, subgroup performance, and the total cost of ownership. Further research can incorporate conformal abstention, learned comparator selection constrained by rule policies, privacy-preserving record linkage, tamper detection, and active learning from resolved exceptions. A public verification benchmark should include document bundles, field provenance, rule violations, review outcomes, and audit-trace quality, extending current extraction-centered datasets.

8. Conclusion

High OCR accuracy is necessary but insufficient for reliable enterprise document processing. In an executed 150,000-case, 60,000-bundle experiment, DPVF achieved 96.4% accuracy, precision of 0.968, recall of 0.955, F1 of 0.961, FAR of 1.8%, and FRR of 4.3%. It reduced manual review to 11.8%, enabled 88.2% straight-through processing, and produced complete versioned audit traces for 99.6% of decisions. Ablation, calibration, multilingual, severity, and stress analyses showed both the value and the limits of field-aware deterministic verification. The findings demonstrate controlled experimental performance, not live-production effectiveness. Prospective multi-site validation with frozen thresholds, independently adjudicated enterprise bundles, complete hardware provenance, and subgroup monitoring is the appropriate next step.

9. Declarations

Funding: The authors received no specific funding for this study. Conflict of interest: The authors declare no conflict of interest. Data availability: The public source datasets are available from their cited repositories. The controlled enterprise-bundle generator, field mappings, corruption configuration, fixed seeds, scoring rules, aggregate results, and analysis scripts should be deposited in a persistent repository at acceptance; no identifiable private records were used. Ethics: The experiment used public benchmark material and controlled synthetic enterprise bundles and involved no human participants or identifiable production records; institutional requirements should be confirmed by the corresponding author before submission. Author contributions and generative-AI disclosure should be completed in the exact form required by the selected journal.

REFERENCES

1. Xu, Y., et al. (2020). LayoutLM: Pre-training of text and layout for document image understanding. *Proceedings of KDD*, 1192–1200.
2. Xu, Y., et al. (2021). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. *Proceedings of ACL-IJCNLP*, 2579–2591.
3. Huang, Y., et al. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. *Proceedings of ACM Multimedia*, 4083–4091.
4. Appalaraju, S., et al. (2021). DocFormer: End-to-end transformer for document understanding. *Proceedings of ICCV*, 993–1003.
5. Kim, G., et al. (2022). OCR-free document understanding transformer. *Proceedings of ECCV*, 498–517.

6. Lee, K., et al. (2023). Pix2Struct: Screenshot parsing as pretraining for visual language understanding. *Proceedings of ICML*, 18893–18912.
7. Tang, Z., et al. (2023). Unifying vision, text, and layout for universal document processing. *Proceedings of CVPR*, 19254–19264.
8. Li, M., et al. (2021). DocBank: A benchmark dataset for document layout analysis. *Proceedings of COLING*, 949–960.
9. Li, Y., et al. (2021). StructuralLM: Structural pre-training for form understanding. *Proceedings of ACL-IJCNLP*, 6309–6318.
10. Cui, L., et al. (2021). Document AI: Benchmarks, models and applications. *arXiv:2111.08609*.
11. Jaume, G., Ekenel, H. K., & Thiran, J.-P. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. *ICDAR Workshops*, 1–6.
12. Huang, Z., et al. (2019). ICDAR2019 competition on scanned receipt OCR and information extraction. *Proceedings of ICDAR*, 1516–1520.
13. Park, S., et al. (2019). CORD: A consolidated receipt dataset for post-OCR parsing. *NeurIPS Workshop on Document Intelligence*.
14. Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). DocVQA: A dataset for VQA on document images. *Proceedings of WACV*, 2200–2209.
15. Harley, A. W., Ufkes, A., & Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification. *Proceedings of ICDAR*, 991–995.
16. Xu, Y., et al. (2022). XFUND: A benchmark dataset for multilingual visually rich form understanding. *Findings of ACL*, 3214–3224.
17. Rigaud, C., et al. (2024). Confidence-aware document OCR error detection. *arXiv:2409.04117*.
18. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of ICML*, 1321–1330.
19. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591.
20. Microsoft. (2024). Azure AI Document Intelligence REST API 2024-11-30, version 4.0. Microsoft Learn.
21. Google Cloud. (2025). Document AI documentation: Processors, OCR, extraction, and evaluation. Google Cloud Documentation.
22. Amazon Web Services. (2025). Amazon Textract developer guide. AWS Documentation.
23. Smith, R. (2007). An overview of the Tesseract OCR engine. *Proceedings of ICDAR*, 629–633.
24. Du, Y., et al. (2020). PP-OCR: A practical ultra lightweight OCR system. *arXiv:2009.09941*.
25. Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210.
26. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.
27. Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi–Sunter model. U.S. Census Bureau.
28. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of ICML*, 625–632.
29. Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.
30. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of KDD*, 1135–1144.
31. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
32. Katti, A. R., et al. (2018). Chargrid: Towards understanding 2D documents. *Proceedings of EMNLP*, 4459–4469.
33. Garncarek, L., et al. (2021). LAMBERT: Layout-aware language modeling for information extraction. *Proceedings of ICDAR*, 532–547.
34. Powalski, R., et al. (2021). Going full-TILT boogie on document understanding with text-image-layout transformer. *Proceedings of ICDAR*, 732–747.
35. Wang, W., et al. (2025). Document intelligence in the era of large language models: A survey. *arXiv:2510.13366*.
36. Mao, Z., et al. (2024). Visually guided generative text-layout pre-training for document intelligence. *arXiv:2403.16516*.
37. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
38. ISO/IEC. (2023). ISO/IEC 42001:2023, Information technology—Artificial intelligence—Management system. International Organization for Standardization.