Automating business document processing: a comparative review of OCR, machine learning and large language models

W. SAKOWSKA
ORCID: 0009-0002-9104-9231

sakowskawiktoria@gmail.com

University of Gdańsk, Faculty of Management, Gdańsk, Poland

Automating business document processing requires more than character recognition: fields must be identified, layout relations preserved and the structured output validated. This paper reviews four families of processing pipelines: optical character recognition (OCR) combined with rules, layout-aware machine learning, OCR--assisted large language models (LLMs) and multimodal models working directly on page images. They are compared in terms of extraction accuracy and reliability, robustness to layout and schema changes, and implementation requirements. On this basis, a qualitative decision framework is proposed that links document structure, layout variability, availability of labelled data, accuracy requirements and operational limitations to a justified starting point for testing. The reviewed evidence indicates that effectiveness depends on the task and on the configuration of the whole pipeline, and that OCR accuracy alone does not determine extraction quality. As no approach is universally best, task-specific evaluation with explicit validation is recommended.

Keywords: intelligent document processing, information extraction, OCR, layout-aware models, multimodal large language models.

DOI: 10.5604/01.3001.0055.9906

1. Introduction

Invoices, purchase orders, receipts and forms convey information through both language and layout. Business process automation typically requires transferring supplier names, dates, order numbers, net, tax and gross amounts, and line items into fixed enterprise records; because manual entry is slow and error-prone, this work is frequently automated. Optical character recognition (OCR) converts image regions into text, yet recognition alone does not determine which value belongs to which field. A single invoice may contain several dates or tax amounts, and the role of each depends on nearby labels, table relations and the entire document structure [6, 8].

The pipelines used for processing documents have changed a great deal. In the traditional approach, OCR is combined with templates and rules; in learned systems, recognised tokens are linked with their coordinates; and generative models create structured fields either from OCR text or directly from page images. These technologies are not entirely distinct from one another, since modern OCR can itself be neural [5], and both layout-aware models and OCR-assisted LLMs still rely on recognition quality [2, 10, 11, 12].

This review asks under which technical and functional conditions different document information extraction methods are most suitable for business process automation. The working assumption is that no single approach is universally best and that suitability depends on document structure, layout variability, data availability, required accuracy and operational limitations. The contribution is a qualitative decision framework linking four pipeline families to these conditions, together with their typical failure modes and validation requirements. Evidence from published comparisons is kept separate from the author's deployment-oriented interpretation. Section 2 describes the review method, Section 3 the four pipeline families, Section 4 compares them, Section 5 presents the framework, Section 6 discusses evaluation in business settings, and Sections 7 and 8 contain the discussion and conclusions.

2. Review method and scope

This paper is a focused narrative review rather than a systematic review or a meta-analysis. Searches were carried out in the ACL Anthology, ACM Digital Library, IEEE Xplore and SpringerLink, while arXiv and Google Scholar were used to find peer-reviewed versions of

preprints and to check metadata. The queries combined terms relating to documents (such as invoice, receipt, purchase order, form and business document) with terms relating to methods (including key information extraction, OCR, layout-aware, OCR-free and multimodal large language model). The searches were performed iteratively in 2026; since this is a narrative review, the number of records found is not reported.

The review included peer-reviewed studies published between 2019 and 2026 that dealt with the extraction of structured information from visually complex or business documents and either proposed an influential architecture, presented a business-document benchmark, or directly compared pipeline configurations. Studies focused solely on handwriting recognition or document classification, as well as non-peer-reviewed vendor material, were excluded. The final collection consists of architecture papers [2, 4, 5, 10, 11, 12], benchmarks [3, 8] and pipeline-level comparisons [1, 7, 9]. The systematic review by Rombach and Fettke [6], which covers 130 deep learning methods published between 2017 and 2024, serves as background; in contrast to that review, the present paper does not classify individual models but examines pipeline families from the point of view of deployment decisions.

Five criteria were applied: (C1) the representation available to the extraction component (text, coordinates or pixels); (C2) the adaptation mechanism (rules, labelled examples, or prompts combined with output schemas); (C3) reported extraction accuracy together with its evaluation context; (C4) robustness to layout changes and unseen templates; and (C5) implementation and operational requirements, such as external dependencies, computing resources, auditability and data handling. C1 and C2 define the families described in Section 3; C3–C5 structure the comparison in Section 4. Comparative statements refer only to the experimental conditions reported by the original authors, because results obtained on different datasets, annotation schemes and metrics cannot be pooled.

3. Document processing approaches

3.1. OCR and rule-based extraction

In an OCR-and-rules pipeline, a recognition engine supplies text and, where available, positions; extraction then relies on patterns, label dictionaries, templates or deterministic checks. For a fixed-layout invoice, a rule may search for a tax identifier near a known label and check its format, or locate an order number with a regular expression. The logic is transparent, and it is clear why a value was selected. The main drawback is maintenance: new suppliers, templates or reading orders require new or modified rules. The approach nevertheless remains a useful baseline even when a learned model is ultimately deployed.

"Traditional" should not be equated with "non-neural": TrOCR, for example, uses pre-trained Transformer models for text recognition [5]. OCR quality must also be distinguished from downstream extraction quality. In an industrial study, Sun et al. [9] show that high character-level OCR scores do not dependably predict downstream performance, because structural and semantic errors can still degrade an OCR-based pipeline even when character error rates are low. A similar error propagation is a plausible concern for field extraction, although its extent must be assessed for each task.

3.2. Layout-aware machine learning

Layout-aware models use words together with their spatial positions, and sometimes visual features, to predict entities or relations. LayoutLM [11] introduced joint pre-training of text and layout; LayoutLMv2 [12] improved cross-modal interaction through visual signals and spatially aware attention; LayoutLMv3 [2] applied unified text and image masking. For LayoutLMv2, the entity labelling score on FUNSD (Form Understanding in Noisy Scanned Documents) rose from 0.7895 to 0.8420 compared with the best previously reported result [12]. This is an improvement under a specific benchmark and protocol, not an indicator of general business benefit.

These models are able to learn repeated position-dependent patterns without the need for a separate rule for each label. Generally, they require recognised tokens, bounding boxes and labelled examples for a particular schema, and any change in annotation labels or the introduction of new supplier formats usually means that further training is needed. The DocILE (Document Information Localization and Extraction) benchmark, which includes 6.7k annotated business documents, specifically addresses the localisation of key information and the recognition of line items, and for this reason reflects invoice processing requirements more accurately than text recognition scores [8].

3.3. OCR-assisted large language models

An LLM can receive OCR output serialised as lines, blocks or key–value pairs and be instructed to produce structured records. Recognition is thus separated from interpretation, so schemas or instructions can be changed without retraining the recogniser. Spatial signals may, however, be lost unless the input contains coordinates, reading order or layout markers; DocLLM shows that generative language models can exploit spatial information for document understanding [10].

Pipeline configuration matters considerably. Colakoglu, Solmaz and Fürst [1] analyse input structuring, model interaction and output refinement for layout-rich extraction. With an optimised configuration they obtained F1 gains of 13.3 to 37.5 points over a common baseline configuration of the same LLM, and concluded that well-configured general-purpose LLMs can match specialised models. Replacing a trained extractor with a prompt therefore does not remove engineering work: input representation, chunking and schema validation remain part of the system. Generated fields also require provenance and consistency checks, since well-formed JavaScript Object Notation (JSON) output is not evidence of correctness.

3.4. OCR-free and multimodal models

OCR-free document understanding maps a page image directly to a target sequence without a separate OCR front end. Donut, an OCR-free document understanding Transformer [4], is an encoder–decoder model designed for specific tasks rather than a general conversational model. Nowadays, multimodal large language models are able to accept images directly and can complete extraction schemas, sometimes using only a small number of task-specific examples.

In a large-scale industrial study on real business documents, Shen et al. [7] found that image-only multimodal extraction can perform comparably to OCR-enhanced methods, and that carefully designed schemas, examples and instructions further improved results. This does not make OCR obsolete, since performance depends on the model, document complexity and schema. Moreover, “OCR-free” does not mean that no text recognition takes place, only that no separate, externally accessible OCR step is required. Image-based systems should therefore be compared with OCR-assisted pipelines on the same documents and use cases.

4. Comparative analysis

Table 1 summarises the four families according to criteria C1–C5. They are treated as configurations rather than fixed model classes, because hybrid solutions are common and a single model may internally perform several steps. Labels such as “low” or “high” cost are deliberately omitted, as hardware, pricing, document length, pre-processing and verification policies differ between deployments.

Tab. 1. Qualitative comparison of document processing approaches

Approach	Input to the extractor (C1)	Adaptation (C2)	Main strength	Main limitation / typical errors
OCR + rules	OCR text, optionally positions	Templates, patterns, label dictionaries	Transparent, auditable extraction logic	Limited flexibility for changing layouts; missed fields when patterns fail [5, 9]
Layout-aware ML	OCR tokens with bounding boxes; optionally image features	Labelled examples; fine-tuning	Learns relations between fields and positions [2, 11, 12]	Requires annotated data; field confusion on unseen layouts [6, 8]
OCR + LLM	Serialised OCR text, optionally with layout markers	Instructions, examples, output schema	Flexible schemas without retraining [1, 10]	Depends on OCR and configuration; unsupported values may be generated [1, 9]
Image-based model	Page image	Fine-tuning [4] or instructions and examples [7]	Uses visual context directly; no separate OCR step [4, 7]	Model- and task-dependent accuracy; outputs require verification [7]

Source: author’s own elaboration based on the literature review

4.1. Accuracy and reliability

The reviewed studies do not support a single accuracy ranking. Layout-aware pre-training improved results on form understanding benchmarks [2, 11, 12]; well-configured LLM pipelines matched specialised models in the settings studied by Colakoglu, Solmaz and Fürst [1]; and image-only multimodal extraction was on par with OCR-enhanced pipelines in the study by Shen et al. [7]. As each result is tied to particular datasets, field definitions and metrics, F1 or accuracy values from different papers cannot be compared directly.

Reliability is better described in terms of failure modes than overall scores. With rule-based extraction, failures can be detected clearly when a pattern does not match, but incorrect values may also be returned when a rule matches content that was not intended. Layout-aware taggers can assign the right token to the wrong field, for instance by mistaking net amounts for gross amounts. Generative models, whether textual or multimodal, might generate well--formed values that are not actually supported by the document. Pipelines that begin with OCR pass both recognition and structural errors downstream [9]. In business environments these differences are more important than minor differences in scores, since they affect which errors can be picked up by validation.

4.2. Flexibility and adaptability

The approaches differ in where domain knowledge is encoded. In rule-based pipelines it resides in templates and patterns, so each new layout requires new or modified rules. Layout-aware models encode it in labelled examples; they generalise to layouts similar to the training data, but new fields or substantially different documents call for more annotation and retraining. Benchmarks such as DocILE make it possible to measure generalisation, as they include line items and diverse layouts [8]. In LLM-based pipelines, knowledge is encoded in instructions, examples and output schemas, which in principle makes schema changes inexpensive [1, 7]. The drawback is sensitivity to prompt and input formatting; moreover, model version updates may change outputs even when the pipeline code is unchanged.

4.3. Computational and implementation requirements

The families also differ in the components that have to be built and maintained. OCR with rules requires an OCR engine, a rule set and validation logic. Layout-aware learning requires an annotation process, training infrastructure and model versioning. OCR-assisted LLMs need inference capacity or access to an external service, together with input serialisation, prompt management and output validation; image-based models omit the OCR front end, but their inference requirements are affected by image resolution and page count. There is also a cost associated with configuration search: Colakoglu, Solmaz and Fürst [1] found that a one-factor-at-a-time search came within 0.8–1.8 F1 points of an exhaustive search while using about 2.8% of its computation. Lastly, when documents contain personal or commercially sensitive data, requirements relating to processing location, retention and access control may eliminate some deployment options even if benchmark results are favourable.

5. Decision framework

Table 2 presents the author's decision framework. It links the conditions named in the working assumption – document structure, layout variability, availability of labelled data, accuracy requirements and operational limitations – to the approach that should be tested first. The recommendations are conditional starting points derived from the literature rather than results of new experiments, and each must be verified on a document-specific validation set.

The framework is meant to be applied iteratively. An organisation first characterises its document stream (number of suppliers, layout stability, share of scanned documents, languages), the available historical data and any data handling constraints. The candidate design is then compared with a simpler baseline, usually OCR with rules, on the same validation set. A more complex method should be adopted only if it reduces field errors or manual review workload on that set, not because it scores higher on a public benchmark.

Tab. 2. Decision framework for business document processing

Business scenario	Candidate starting point	Rationale	Validation priority
Recurring invoices from few suppliers; stable layouts	OCR + rules	Explicit, auditable logic suffices; low variability limits rule maintenance	Field errors after template changes; onboarding of new suppliers
Many layouts; labelled historical fields available	Layout-aware ML	Learns spatial patterns from examples [2, 11, 12]	Generalisation to unseen layouts and line items [8]
Variable or frequently changing schemas; reliable OCR	OCR-assisted LLM	Schema changes via instructions; configuration strongly affects results [1]	Grounding of values in source text; schema compliance
Irregular visual layouts; difficult OCR integration	Image-based multimodal model	Competitive with OCR-enhanced pipelines under reported conditions [7]	Exact values, tables and cross-page consistency
High-value fields with strict arithmetic or legal constraints	Any of the above with deterministic reconciliation	Business rules catch errors that extraction models do not report	Consistency of totals, taxes and identifiers; share routed to manual review

Source: author's own elaboration based on the literature review

Hybrid solutions are explicitly allowed – for example OCR with a trained field tagger and deterministic reconciliation, or direct image extraction followed by rule-based checks – combining the flexibility of learned models with the predictability of rules.

6. Evaluation in business settings

A fair comparison requires fixed documents, field definitions and acceptance criteria. The evaluation set should consist of both known and new layouts, low-quality scans and cases involving missing or duplicated field labels. FUNSD includes 199 annotated forms [3]; DocILE contains 6.7k annotated business documents and explicitly assesses line items [8]. The results obtained from these benchmarks cannot be combined unless differences in annotation, data and metrics are taken into account.

For each target field, exact-match accuracy or accuracy on normalised numbers and dates should be reported separately from token-level F1. When evaluating line items, attention should be paid to the relationship between items and their values rather than to individual tokens. Assessment should also include checks for missing fields, hallucinated values, evidence localisation, arithmetic consistency (for example, whether the net amount and the tax add up to the gross amount), exception handling and the proportion of documents that need manual review. Although OCR character or word error rates can be helpful for diagnosis, they are not sufficient as a final business measure, as Sun et al. [9] demonstrated for a related downstream task. A credible operational comparison should also provide information on processing time, computing or service fees, document handling constraints and the human effort required for annotation, maintenance or rule updates; the present review proposes these criteria but does not assign relative importance to them.

7. Discussion and limitations

The evidence examined supports three observations. Firstly, spatial relationships matter: layout-aware models are able to disambiguate fields that cannot be determined from the text alone [2, 10, 11, 12]. Secondly, the quality of LLM-based extraction depends heavily on the entire pipeline, covering both input representation and output refinement [1]. Thirdly, the possibility of skipping a separate OCR stage is an empirical rather than a universal finding: Shen et al. [7] found direct image extraction competitive under the conditions assessed, whereas Sun et al. [9] demonstrated that OCR-based pipelines can fail at structural interpretation even if recognition appears satisfactory.

Several trade-offs are involved. While instruction-driven methods are easy to adapt to new schemas, they tend to make less predictable errors than rule-based methods. Rules can be put in place quickly for a small number of templates but become expensive to maintain as suppliers

change, whereas learned models shift the cost onto annotation and retraining. Multimodal models reduce rather than remove the dependence on recognition, and in all cases validation and monitoring of generalisation to new layouts are still required. These observations suggest a layered design in which recognition, semantic extraction, schema validation and exception review are each assessed separately, and business rules – such as currency normalisation, arithmetic checks and supplier identifiers – are kept outside the model.

The review has limitations. Source selection is purposive rather than exhaustive; the included studies use different data, metrics and model versions; and the decision criteria are qualitative and have not been tested in a controlled setting. Published experiments do not establish production costs, privacy compliance or error rates for any particular organisation. Since the field develops quickly, conclusions about specific multimodal models will become outdated sooner than those about pipeline structure.

8. Conclusions

In answer to the research question, the suitability of a document extraction method should be judged by the conditions of its use rather than by its architecture alone. OCR with rules is a reasonable starting point for stable layouts and strict auditability; layout-aware learning suits diverse layouts with labelled history; OCR-assisted LLMs suit changing schemas, provided that OCR is reliable; and image-based multimodal models deserve testing for visually irregular documents. Business document automation should therefore be treated as the design of an extraction and validation workflow rather than as the choice of a single model. Future work needs to benchmark the four pipeline families on a shared, versioned set of invoices and purchase orders, reporting field-level accuracy, hallucinated values and manual review effort, with particular attention to unseen layouts and multilingual documents.

9. Bibliography

[1] Colakoglu G., Solmaz G., Fürst J., "Problem Solved? Information Extraction Design Space for Layout-Rich Documents using LLMs", *Findings of the Association for Computational Linguistics: EMNLP 2025*, China, Suzhou 2025, pp. 17908–17927, Association for Computational Linguistics, 2025 https://doi.org/10.18653/v1/2025.findings-emnlp.973.

[2] Huang Y., Lv T., Cui L., et al., "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking", *Proceedings of the 30th ACM International Conference on Multimedia*, Lisboa 2022, pp. 4083–4091, Association for Computing Machinery, New York, United States, 2022, https://doi.org/10.1145/3503161.3548112.

[3] Jaume G., Ekenel H.K., Thiran J.-P., "FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents", *2019 International Conference on Document Analysis and Recognition Workshops* (*ICDARW*), Sydney 2019, https://doi.org/10.1109/ICDARW.2019.10029.

[4] Kim G., Hong T., Yim M. et al., "OCR-free Document Understanding Transforme", *Computer Vision – ECCV 2022*, LNCS, 13688, 498–517, Springer, Cham 2022, https://doi.org/10.1007/978-3-031-19815-1_29.

[5] Li M., Lv T., Chen J., et al., "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 11, 13094–13102 (2023), https://doi.org/10.1609/aaai.v37i11.26538.

[6] Rombach A.M., Fettke P., "Deep Learning based Key Information Extraction from Business Documents: Systematic Literature Review", *ACM Computing Surveys*, Vol. 58, No. 2, 1–37 (2025), https://doi.org/10.1145/3749369.

[7] Shen J., Peiyue Y., Ghosh A., et al., "OCR or Not? Rethinking Document Information Extraction in the MLLMs Era with Real-World Large-Scale Datasets", *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 5: Industry Track, pp. 385–396, Association for Computational Linguistics, Rabat 2026, https://doi.org/10.18653/v1/2026.eacl-industry.28.

[8] Šimsa Š., Šulc M., Uřičář M., et al., "DocILE Benchmark for Document Information Localization and Extraction", *Document Analysis and Recognition – ICDAR 2023*, LNCS 14188, pp. 147–166, Springer, Cham 2023, https://doi.org/10.1007/978-3-031-41679-8_9.

[9] Sun L., Wangdexian, Huang J., et al., "When Good OCR Is Not Enough: Benchmarking OCR Robustness for Retrieval-Augmented Generation", *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, San Diego 2026, Vol. 6: Industry Track, pp. 884–894, Association for Computational Linguistics, 2026, https://doi.org/10.18653/v1/2026.acl-industry.60.

[10] Wang D., Raman N., Sibue M., et al., "DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding", *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok 2024, Vol. 1: Long Papers, pp. 8529–8548, Association for Computational Linguistics, 2024, https://doi.org/10.18653/v1/2024.acl-long.463.

[11] Xu Y., Li M., Cui L., et al., "LayoutLM: Pre-training of Text and Layout for Document Image Understanding", *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1192–1200, Association for Computing Machinery, NY, United States 2020, https://doi.org/10.1145/3394486.3403172.

[12] Xu Y., Xu Y., Lv T., et al., "LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1: Long Papers, 2579–2591, Association for Computational Linguistics, 2021, https://doi.org/10.18653/v1/2021.acl-long.201.

Automatyzacja przetwarzania dokumentów biznesowych: przegląd porównawczy OCR, uczenia maszynowego i dużych modeli językowych

W. SAKOWSKA

Automatyzacja przetwarzania dokumentów biznesowych wymaga rozpoznawania znaków, identyfikacji pól, zachowania relacji przestrzennych oraz sprawdzania poprawności danych wyjściowych. W artykule przedstawiono przegląd literatury dotyczący czterech typów rozwiązań: OCR opartego na regułach, modeli uwzględniających układ dokumentu, dużych modeli językowych wykorzystujących OCR oraz modeli multimodalnych analizujących obraz bezpośrednio. Porównano je pod kątem dokładności i niezawodności ekstrakcji, odporności na zmiany układu i schematu danych oraz wymagań wdrożeniowych. Na tej podstawie opracowano jakościowe ramy decyzyjne, które łączą strukturę dokumentów, zmienność układu, dostępność oznaczonych danych, wymagania dotyczące dokładności oraz ograniczenia organizacyjne z wyborem podejścia do testowania. Analiza pokazuje, że skuteczność zależy od zadania i konfiguracji całego procesu, a sama dokładność OCR nie pozwala ocenić jakości ekstrakcji informacji. Nie wskazano rozwiązania najlepszego dla każdego przypadku; zaproponowano ocenę w konkretnych warunkach użytkowania.

Słowa kluczowe: inteligentne przetwarzanie dokumentów, ekstrakcja informacji, OCR, modele uwzględniające układ, multimodalne modele językowe.