Metric Blindness in Document Information Extraction

Why character-level accuracy cannot certify structured extraction from high-stakes documents

Kuluru Vineeth

kuluruvineeth8623@gmail.com

August 2026

doi:10.5281/zenodo.21755329

Abstract

Vision-language models now outperform traditional optical character recognition on the character- and word-level metrics by which document extraction systems are ranked. We argue that this ranking is not merely imprecise but structurally uninformative for high-stakes extraction, because those metrics fail *selectively*: they register the loud failures of generative readers while remaining close to blind to the most consequential one, fluent semantic substitution of named entities and numeric values on legible input. We synthesise evidence across three architectural lineages and identify a three-layer measurement failure: character-level scores are insensitive to entity substitution, established line-item metrics have been judged misaligned with practical use by the authors of the systems that lead them, and schema-validity checks certify structure while saying nothing about value correctness. We further observe that the public benchmark landscape terminates before the document class in which these failures are most consequential: no peer-reviewed benchmark evaluates extraction from tax returns or comparable regulatory filings, and none evaluates cross-field arithmetic consistency in any document class. We argue for replacing point-accuracy reporting with calibrated coverage reporting, for which conformal prediction over key-information extraction supplies a demonstrated mechanism, and we specify the benchmark whose absence currently makes accuracy claims in this domain unfalsifiable. This is a position and synthesis paper; we report no new experiments.

1 Introduction

A practitioner selecting a document extraction system in 2026 encounters a consistent claim: field-level accuracy between 97% and 99%. The claim is rarely accompanied by a method, and the literature gives strong reason to believe that where a method exists, it measures something other than what the practitioner needs to know.

The immediate evidence is a benchmark of seventeen traditional OCR and vision-language systems on 175 microfilm scans of archival documents [4]. Vision-language models swept median word error rate and took the best reported figure on every statistic, though the margin is narrower than the framing suggests: the strongest vision model improves mean character error rate (CER) over the best traditional system by 1.36 points, and several vision models score worse. A qualitative analysis of the same outputs then found systematic failure modes invisible to those metrics: orthographic normalisation, spurious content generation, and semantic substitutions that preserve fluency while altering meaning. Errors affecting named entities were singled out as particularly critical, since they introduce substantial semantic distortion at minimal metric cost.

Three details from that study carry the argument further than the headline. The leading system records a mean CER of 11.17% against a median of 1.72%—a sixfold gap indicating that an aggregate score summarises a heavily skewed distribution in which occasional catastrophic outputs are averaged into apparent competence. The semantic substitutions occur predominantly on *clearly legible* regions, and are attributed to the decoder deviating from source content under learned entity priors rather than to image degradation, which is what allows the finding to travel beyond damaged microfilm. And the substitutions persist when models are explicitly instructed to mark unreadable text as illegible—prompted abstention does not suppress them.

This is therefore not a marginal observation about a hard corpus. It is a statement about the relationship between a metric and an error distribution, and it generalises to any setting where a generative model transcribes values that are consumed downstream as facts. A tax identification number, a subtotal, or a jurisdiction code that has been replaced with a well-formed alternative is a defect that the measuring instrument is constructed not to see.

That aggregate character metrics understate downstream harm is itself established rather than novel: OCR errors affecting named entities degrade recognition and linking disproportionately to their character cost [6], and the effect persists in retrieval [9]. The closest prior position to ours argues that character and word error rates fail to capture downstream impact specifically for business-document key-field and line-item extraction [18], and a concurrent large-scale study revisits whether OCR belongs in the pipeline at all under multimodal models [21]. We take that critique as established and extend it in three directions those works do not address: to schema-validated structured output, to cross-field arithmetic consistency, and to calibrated coverage as a reporting unit for regulatory filings.

We make three claims. First, that metric blindness is a property of the metric–architecture pair rather than an arte-

Lineage	Guarantee	Cost
Layout-aware discriminative	Cannot emit a value absent from the page	Supervision required per document type
OCR-free generative	No OCR error propagation	Substitution surface opens
LLM-centric	Generalises across formats; infers, reasons	Full generative risk; layout read only via scaffolding

Table 1: The lineages differ less in accuracy than in the class of error each can commit. The guarantee in row one was traded away, not superseded.

fact of any particular dataset (§3). Second, that the natural remedies practitioners reach for—structured-output constraints and schema validation—address a different failure entirely (§4). Third, that the benchmark landscape has a specific and consequential hole, and that filling it is tractable (§6). We close by arguing for calibrated coverage as the reporting unit (§7) and specifying a benchmark (§8).

2 Three lineages, three failure modes

The architectural history of document extraction is usually narrated as a progression in accuracy. It is more usefully read as a progression in *what kind of error each system is capable of committing*, summarised in Table 1.

The discriminative layout-aware lineage began with joint pre-training over text and two-dimensional position [28], extended to multimodal pre-training incorporating image features [27], and diversified into variants that discard the vision encoder while retaining spatial reasoning [7] or decouple layout from language to enable cross-lingual transfer [26, 29]. These models assign labels to tokens that exist on the page. The property this confers is not a performance characteristic but a structural one: a token classifier cannot emit a value that is absent from its input. Its cost is equally structural—supervision is required per document type, and performance degrades on unseen layouts.

The generative OCR-free lineage removed the recognition stage entirely, mapping document images to structured output through an image encoder and text decoder [13], later unified across vision, text and layout in a single generative interface [24]. This eliminated OCR error propagation and simultaneously opened the substitution surface, because a decoder that produces the answer can produce an answer with no support in the image.

The current lineage places a language model at the centre, teaching layout either through instruction tuning [15] or through disentangled spatial attention over bounding boxes [25]. Approximately 300 papers published between 2021 and mid-2025 fall under this heading [12]. It inherits the generative failure surface in full.

2.1 Scale is not the remedy

A result that bears directly on system design is that the largest multimodal model is frequently the wrong investment. Reframing business document extraction as a tool-use problem—where the tools are the downstream systems consuming the output—and pairing an ordinary text language model with retrieval-augmented structured generation achieved state-of-the-art results on both key-information extraction and line-item recognition, matching or exceeding large multimodal models lacking the retrieval scaffold [2]. The authors conclude that text models with scaffolding are typically superior under real-world constraints, the multimodal advantage being marginal where present.

An independent line of work converges from the opposite direction, reporting that generative language models struggle to comprehend layout cues in previously unseen formats, and that segmentation quality rather than model scale is the binding bottleneck: supplied with ground-truth segments, a 7B model recovers most of the gap to far larger ones [1]. Both results direct engineering effort toward the retrieval and segmentation layer.

That same study complicates the lineage picture in a way worth stating against our own interest. On its cross-format transfer test, an unaugmented frontier language model scored 95.39 where a discriminative layout model collapsed to 33.43, and segmentation added under two further points [1]. Weak layout reading in generative models is therefore a weakness relative to their own potential, not relative to the discriminative lineage, which degrades far more sharply off-distribution. The trade in Table 1 is between a structural guarantee and format robustness—not between accuracy and inaccuracy.

Generalisation nonetheless remains the binding constraint regardless of lineage, and the magnitude is striking. Across fifteen state-of-the-art large multimodal models on a unified key-information extraction benchmark of 6,133 documents, per-scenario field-level F1 varied by up to 45.6 points *within a single model*, and merely loosening the task from a predefined schema to open-category extraction cost one leading model 20.2 points and another 26.7 [11]. This is consistent with the earlier finding that OCR capability within multimodal models is weaker and less uniform than headline benchmark scores suggest [14]. A single averaged benchmark figure conceals a spread of this size by construction.

3 Metric blindness

Consider the information a metric can carry. CER and WER are edit distances over character and token sequences; their magnitude is proportional to the number of symbols altered, not to the semantic weight of the symbols. Substituting one digit in a nine-digit identifier costs approximately $1/9$ of that field's contribution to the score, whatever the consequence of the substitution. The metric is behaving exactly as specified. The specification is the problem.

Figure 1 arranges the common metrics against the error classes they may encounter. Two observations follow. First,

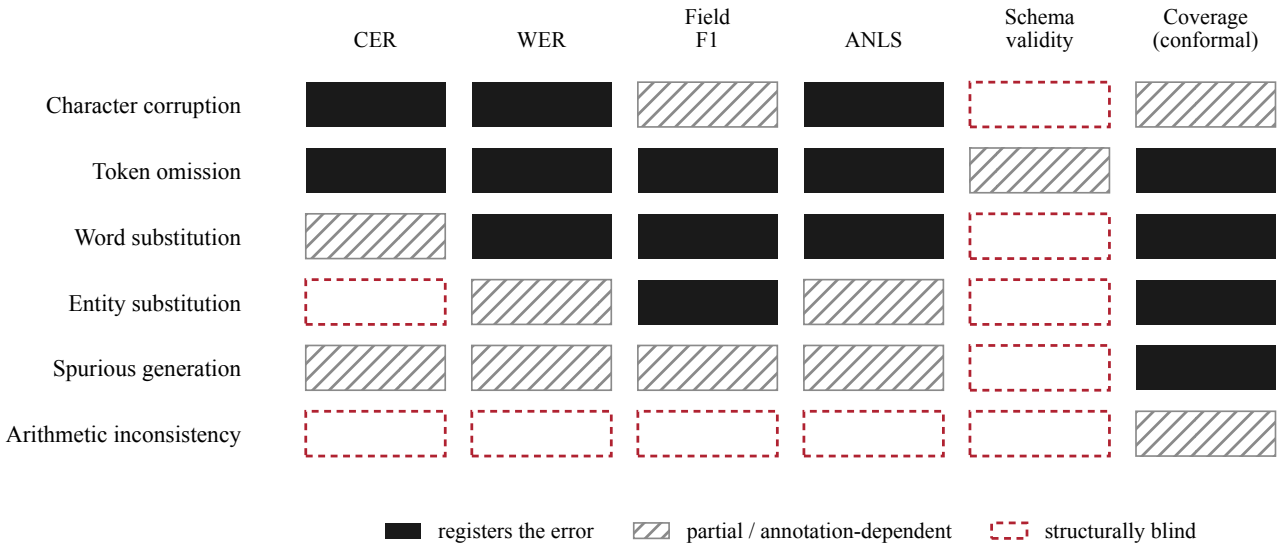

Figure 1: Sensitivity of common evaluation metrics to error classes in document extraction. Entity substitution—singled out as the most consequential failure mode for generative readers [4]—is precisely the class that character-level metrics register weakest, while cross-field arithmetic inconsistency is unregistered by every metric in general use. **This figure is an analytical framework derived from the definitions of each metric, not an empirical measurement; no source reports failure-mode frequencies from which prevalence could be inferred.**

entity substitution—singled out as the most consequential error class for generative readers, since it introduces large semantic distortion while contributing only marginally to aggregate metrics [4]—sits in the region where character-level metrics are weakest, so the metrics that most favour generative systems are the ones least able to detect their characteristic defect. Second, cross-field arithmetic inconsistency is registered by no metric in general use, despite being trivially checkable in any document that contains a total.

Two qualifications keep this precise. Aggregate metrics do not fail uniformly—they fail *selectively*. Generative readers exhibit both a loud failure mode that CER and WER register well (truncated or collapsed transcription) and a quiet one they do not, and it is the selectivity, not a blanket insensitivity, that makes the aggregate misleading. And prevalence is not established: the qualitative taxonomy we rely on reports no failure-mode frequencies, so the claim available from the literature concerns the *consequence* of entity substitution rather than its rate.

The consequence is that improvements in reported accuracy and increases in deployed risk are not merely compatible; under a shift from discriminative to generative architectures we argue they are the expected joint outcome. We note this is our synthesis: no source we surveyed measures a correlation between aggregate score and substitution rate, because none counts substitutions.

3.1 The established metrics are contested by their own users

This is not an outsider's objection. The authors who set the state of the art on line-item recognition simultaneously proposed a replacement metric class, on the grounds that the existing measures were insufficiently aligned with practical extraction use cases [2]. When the group producing the leading numbers publishes a critique of the scoreboard in the same paper, the scoreboard should be treated as provisional.

4 The structured-output illusion

The remedy most readily available to practitioners is constrained decoding against a schema, and it is widely assumed to bound the failure. It fails to, for two independent reasons.

The first is that the guarantee is narrower than advertised even on its own terms. A rigorous evaluation of six constrained-decoding frameworks over 10,000 real-world JSON schemas measured *empirical* coverage—the fraction of schemas a framework declares it supports whose outputs actually validate—and found it far below the declared figure, falling to 0.03 for one framework on hard real-world schemas, with the best framework supporting roughly twice as many schemas as the worst [5]. Schema compliance is therefore not a guarantee a practitioner receives automatically; it is a property that varies by an order of magnitude across implementations of the same nominal feature.

The second reason is the one that matters more here, and we state it as our own analytical claim because no benchmark result establishes it: schema compliance is a statement about the *shape* of an output, whereas the failure under discussion is a statement about its *content*. The two are orthogonal. An object in which every field is present, correctly typed, and factually wrong satisfies every structural validator a practitioner would naturally write, and will do so silently. We stress that the constrained-decoding literature does not claim otherwise—indeed the same evaluation found that constraining generation slightly *improved* downstream task accuracy on reasoning benchmarks—but neither does it measure value correctness on documents, and the assumption that structural

validation transfers to factual validation is supplied by the practitioner rather than by the evidence.

5 Arithmetic as discarded verification

Numerical reasoning over financial documents constitutes a distinct and more severe failure. It was established as its own task by expert-annotated benchmarks over earnings reports and hybrid tabular–textual content [3, 30], remains sufficiently difficult that specialised models continue to be developed rather than relying on general language-model capability [31], and direct evaluation confirms the deficit persists in current systems [23].

The design consequence is that derived quantities should be recomputed deterministically from extracted operands rather than generated. The stronger observation is that documents of this class carry an arithmetic dependency graph—subtotals summing to totals, line items reconciling against balances, prior-period values constraining current ones—which functions as an intrinsic checksum. Delegating computation to the extractor does not merely risk error; it discards a verification signal the document supplies at no cost. We note the absence of this check in Figure 1 as the most readily correctable gap identified in this paper.

6 The benchmark landscape and its hole

Figure 2 maps the public evaluation surface. Receipts [8], noisy forms [10], document-image question answering [17], invoices with line items [22], and financial-report reasoning [3, 30] are all covered. Earlier work established that field extraction generalises across templates when a model learns field representations rather than absolute positions [16], a result about form-like documents in general.

To our knowledge no peer-reviewed benchmark evaluates extraction from tax *returns* or comparable regulatory filings. One distinction must be drawn carefully, since we cite the counterexample ourselves: the unified benchmark discussed above does include a tax-compliant document scenario [11], but those are transactional instruments averaging eight fields per document—tax invoices in the value-added sense—rather than filings. A personal or corporate return carries hundreds of interdependent fields, and nothing in the indexed literature addresses jurisdiction-specific form variation, multi-year carryforward consistency, or the arithmetic dependency graph described above. The gap is in filings, not in tax-adjacent documents generally. Two consequences follow. Accuracy claims in this domain are not currently falsifiable, since no shared evaluation exists against which to falsify them. And the column corresponding to cross-field consistency is empty for *every* document class, not only for tax—the verification signal identified in the preceding section is unmeasured throughout the field.

7 Coverage in place of accuracy

The literature offers one mechanism that converts this problem into a managed quantity. Applying split conformal prediction post hoc to fine-tuned multimodal extractors on a receipt corpus yields entity-level prediction sets satisfying a user-specified error rate: reported marginal coverage was 98.3% at $\alpha = 0.02$, with 70% of predictions returned as high-confidence singletons [20]. Reliability varied by field in an interpretable way—highly structured fields such as dates and prices produced small sets with near-perfect reliability, while rare or semantically ambiguous fields produced large sets and lower coverage—and the calibrated set sizes were used directly to automate confident extractions and route uncertain ones for human review.

We argue this should become the reporting convention rather than an optional addition. A coverage statement—*this system processes x% of fields automatically at a guaranteed error rate α*—is falsifiable, auditable, and carries the per-field structure that a single accuracy figure destroys. A point accuracy figure, by contrast, is a summary over an unstated distribution measured by an instrument shown in §3 to be insensitive to the relevant error class. Comparable architectures are being assembled in adjacent regulated workflows [19], suggesting the convention is not specific to document extraction.

Figure 3 places calibration in a full pipeline consistent with §2: segmentation and retrieval feeding a text model under schema constraint, deterministic recomputation of derived values, and calibrated set size governing the human interface.

8 A benchmark proposal

We specify what the missing benchmark must contain, since the gap in Figure 2 is tractable. It requires: (i) tax or equivalent regulatory filings across at least two jurisdictions and two filing years, to expose the distribution shift that periodic form redesign induces; (ii) field-level ground truth with entity-type annotation, so that substitution in identifiers and amounts is scored separately from character-level noise; (iii) declared arithmetic dependencies between fields, making cross-field consistency directly scorable; (iv) a held-out calibration split, so that coverage rather than accuracy can be reported; and (v) a hallucination probe—documents containing fields that are absent, illegible, or ambiguous, where the correct output is abstention.

Criterion (v) is the one no surveyed benchmark *scores*, and it cannot be satisfied at the prompt. The nearest existing practice measures faithfulness post hoc, checking whether predicted values can be located in the document and classifying those that cannot as hallucinated [11]; that is an analysis correlated against F1, not a task in which abstention is the correct output and is rewarded as such. The distinction matters because semantic substitutions persist in models explicitly instructed to mark unreadable regions as illegible [4], so abstention has to be trained and scored rather than requested. A system that never abstains cannot be distinguished, under current scoring, from one that abstains correctly.

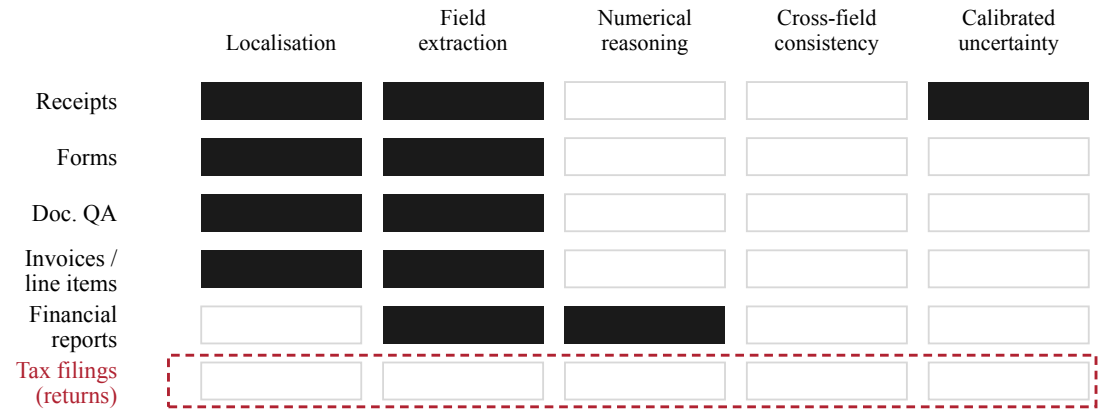

Figure 2: Public benchmark coverage by document class and evaluated property. Calibrated uncertainty has been demonstrated only on receipts [20]; cross-field consistency is evaluated nowhere; and the row for tax *filings* is empty in every column. Tax-adjacent transactional documents (invoices averaging eight fields) do appear in existing benchmarks [11]; multi-field returns do not.

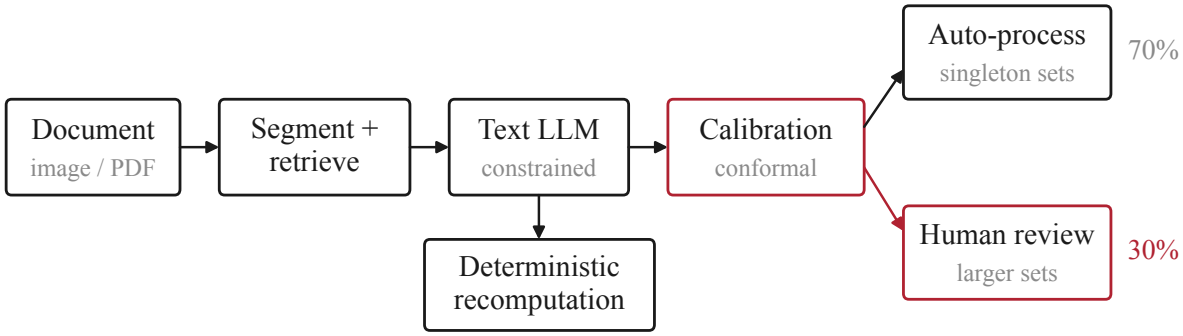

Figure 3: Risk-aware extraction. Calibrated prediction-set size, not a point estimate, determines routing. The 70/30 split is the reported operating point at $\alpha = 0.02$ on receipts [20]; deterministic recomputation supplies an independent check on derived values.

9 Limitations

This is a synthesis and position paper and reports no new experiments; every empirical value cited is attributed to the work that measured it. The four sources on which our argument most depends were read in full [4, 11, 5, 1]; the remainder of the survey was assembled from bibliographic metadata and abstracts retrieved through Crossref and OpenAlex, which is adequate for establishing the shape of a literature and the presence of a gap but insufficient to critique those methodologies individually.

The load-bearing evidence is also narrower than our argument. The study establishing that semantic substitution escapes character metrics evaluates 175 pages of bi-level microfilm of Spanish-language typewritten documents, reports no error bars, includes no frontier commercial systems, and measures full-page *transcription* rather than field extraction; it also reports no failure-mode frequencies, so no prevalence claim is available from it. Its finding that substitutions occur on legible regions under decoder priors, rather than under image degradation, is what licenses our extrapolation to clean business documents—and that extrapolation remains an argument rather than a measurement. No source we cite tests numeric or arithmetic fields at all, so §6 rests on the structure of such documents rather than on evidence about them. Figure 1 is an analytical framework derived from metric definitions, not a measurement, and reasonable disagreement about individual cells is possible. Our claim that no tax-extraction benchmark exists is a negative result over indexed literature and cannot exclude proprietary or unindexed evaluations. Finally, the conformal prediction result we build upon was demonstrated on receipts; whether its guarantees survive the distribution shift of a new filing year is precisely the open question our proposed benchmark is designed to answer.

10 Conclusion

The field ranks document extraction systems using instruments that are least sensitive exactly where generative architectures are most likely to fail, validates their outputs with structural checks that certify shape rather than content, and leaves unmeasured an arithmetic consistency signal that many target documents supply for free. The correction is available: report calibrated coverage rather than point accuracy, recompute derived values deterministically, score entity substitution as its own class, and require abstention to be evaluable. What is missing is a shared benchmark on which such reporting could be compared—and in the docu-

ment class where the stakes are highest, that benchmark does not exist.

References

[1] Aniket Bhattacharyya, Anurag Tripathi, Ujjal Das, et al. Information extraction from visually rich documents using LLM-based organization of documents into independent textual segments. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2505.13535. preprint.

[2] Franz Louis Cesista, Rui L. Aguiar, Jason Z. Kim, et al. Retrieval augmented structured generation: Business document information extraction as tool use. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2024. doi: 10.1109/mipr62202.2024.00042.

[3] Zhiyu Chen, Wenhu Chen, Charese Smiley, et al. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. doi: 10.18653/v1/2021.emnlp-main.300.

[4] Marina Gardella, Camilo Mariño, Diego Belzarena, et al. When low CER is not enough: An analysis of hallucinations in vision-language OCR systems on historical uruguayan documents. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2607.24077. preprint.

[5] Saibo Geng, Hudson Cooper, Michał Moskal, et al. JSONSchemaBench: A rigorous benchmark of structured outputs for language models. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2501.10868. preprint.

[6] Ahmed Hamdi, Elvys Linhares Pontes, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. In-depth analysis of the impact of OCR errors on named entity recognition and linking. *Natural Language Engineering*, 29(2):425–448, 2023. doi: 10.1017/S1351324922000110.

[7] Teakgyu Hong, Donghyun Kim, Mingi Ji, et al. BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. doi: 10.1609/aaai.v36i10.21322.

[8] Zheng Huang, Kai Chen, Jianhua He, et al. ICDAR2019 competition on scanned receipt OCR and information extraction. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2019. doi: 10.1109/icdar.2019.00244.

[9] Alexandre Jaud, Ahmed Hamdi, Antoine Doucet, Adam Jatowt, and Mickaël Coustaty. Beyond CER and WER: How does OCR really impact information retrieval? In *2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 30–39, 2025. doi: 10.1109/JCDL67857.2025.00014.

[10] Guillaume Jaume, Hazım Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019. doi: 10.1109/icdarw.2019.10029.

[11] Yifan Ji, Zhipeng Xu, Zhenghao Liu, et al. UNIKIE-BENCH: Benchmarking large multimodal models for key information extraction in visual documents. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2602.07038. preprint.

[12] Wenjun Ke, Yifan Zheng, Youlan Li, et al. Large language models in document intelligence: A comprehensive survey, recent advances, challenges, and future trends. *ACM Transactions on Information Systems*, 2025. doi: 10.1145/3768156.

[13] Geewook Kim, Teakgyu Hong, Moonbin Yim, et al. OCR-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. doi: 10.1007/978-3-031-19815-1_29.

[14] Yuliang Liu, Zhang Li, Mingxin Huang, et al. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 2024. doi: 10.1007/s11432-024-4235-6.

[15] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, et al. LayoutLLM: Layout instruction tuning with large language models for document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/cvpr52733.2024.01480.

[16] Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, et al. Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.580.

[17] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. doi: 10.1109/wacv48630.2021.00225.

[18] Ngoc Nhi Nguyen, Ahmed Hamdi, Antoine Doucet, Adam Jatowt, and Mickaël Coustaty. Rethinking OCR evaluation for information extraction in business documents. In *Lecture Notes in Computer Science*, pages 245–254, 2026. doi: 10.1007/978-981-95-4861-3_21.

[19] Robert Richardson, Josh Meyers, Brian Hartman, et al. Agentic AI and retrieval-augmented models in straight-through underwriting. *arXiv preprint*, 2026. doi: 10.48550/arXiv.2607.07858. preprint.

[20] Alexander Rombach and Nijat Mehdiyev. Beyond accuracy: Understanding model confidence in key information extraction with conformal prediction. *International Journal on Document Analysis and Recognition (IJDAR)*, 2026. doi: 10.1007/s10032-026-00572-y.

[21] Jiyuan Shen, Peiyue Yuan, Atin Ghosh, et al. OCR or not? rethinking document information extraction in the MLLMs era with real-world large-scale datasets. In *Proceedings of EACL 2026 (Industry Track)*, 2026. doi: 10.48550/arXiv.2603.02789.

[22] Štěpán Šimsa, Milan Šulc, Michal Uřičář, et al. DocILE benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2023. doi: 10.1007/978-3-031-41679-8_9.

[23] Pragya Srivastava, Manuj Malik, Vivek Gupta, et al. Evaluating LLMs' mathematical reasoning in financial document question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.231.

[24] Zineng Tang, Ziyi Yang, Guoxin Wang, et al. Unifying vision, text, and layout for universal document processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/cvpr52729.2023.01845.

[25] Dongsheng Wang, Natraj Raman, Mathieu Sibue, et al. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.463.

[26] Jiapeng Wang, Lianwen Jin, and Kai Ding. LiLT: A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. doi: 10.18653/v1/2022.acl-long.534.

[27] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. doi: 10.18653/v1/2021.acl-long.201.

[28] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. doi: 10.1145/3394486.3403172.

[29] Yiheng Xu, Tengchao Lv, Lei Cui, et al. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. doi: 10.18653/v1/2022.findings-acl.253.

[30] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, et al. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. doi: 10.18653/v1/2021.acl-long.254.

[31] Fengbin Zhu, Ziyang Liu, Fuli Feng, et al. TAT-LLM: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 2024. doi: 10.1145/3677052.3698685.