OCR Large Models: Recent Progress and Prospects

Yi Sun, Zhongheng Zhou, Huiguo He*, Jiahuan Cao, Yuyi Zhang, Peirong Zhang, and Lianwen Jin*

School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China

Abstract. Optical character recognition (OCR) is being redefined by large language models, multimodal large language models, and document-specialized vision-language models. The field is moving from text transcription toward auditable, structured, and machine-actionable document knowledge. This survey reviews recent OCR large models from the perspectives of data, evaluation, architecture, and deployment. We organize representative systems into layout-guided pipelines, end-to-end OCR VLMs, and agentic OCR systems, and discuss high-resolution encoding, visual token compression, and verifiable reinforcement learning. We further summarize data engines, benchmarks, model comparisons, and open challenges in reliability, structure parsing, multilingual robustness, RAG workflows, privacy, latency, and reproducibility.

Keywords: OCR large models · Multimodal large language models · Structured document parsing

1 Introduction

Optical character recognition (OCR) has traditionally been treated as a perception task: given an image, detect text regions, recognize character sequences, and return plain text. This view supported applications in scanned-book digitization, invoice processing, and scene-text understanding. Yet current applications require richer outputs: complex layouts with formulas, tables, figures, multi-column structures, degraded scripts, and handwriting cannot be captured by plain text alone.

The rapid progress of large language models (LLMs) and multimodal large language models (MLLMs) has changed the role of OCR. OCR is now a data engine for LLM pretraining, retrieval-augmented generation (RAG), document question answering, information extraction, and document agents. A useful OCR model must output not only words but also structure, including reading order, boxes, tables, formulas, JSON, Markdown, HTML, LaTeX, semantic labels, and evidence regions. The target has shifted from text recognition to reliable document knowledge extraction. Recent work extends transformer-based OCR and

* Corresponding authors: Huiguo He (hehuiguo@scut.edu.cn) and Lianwen Jin (eelwjin@scut.edu.cn).

OCR-free document models into OCR-specialized VLMs, compact document VLMs, hybrid layout-guided parsers, and tool-augmented agents.

At the same time, general-purpose foundation models have become stronger visual readers, increasingly able to process text-rich images, charts, tables, and documents. However, they are expensive, difficult to audit, sometimes overgenerate, and may not provide stable region-level evidence, motivating OCR-specialized large models that are smaller, more controllable, and trained with document-specific supervision.

This paper surveys recent OCR large models, defined broadly as systems that apply large-scale language, visual, or vision-language modeling to text-rich image recognition, document parsing, structured conversion, or grounded information extraction. This includes end-to-end VLMs, compact OCR VLMs, hybrid layout-guided systems, and production toolkits.

The contributions of this survey are threefold. First, we place recent OCR large models in the broader trajectory of OCR, document intelligence, and multimodal foundation models. Second, we organize representative models into three architectural families: layout-guided pipelines, end-to-end OCR VLMs, and agentic OCR systems, while also discussing common enabling techniques. Third, we identify open problems and future directions, with emphasis on reliability, multilinguality, reading order, visual token efficiency, verifiable rewards, and deployment.

Compared with broader document-intelligence surveys, this work focuses more specifically on OCR large models and their system-level design.

2 Background

2.1 OCR to Document Intelligence

Classical OCR systems usually decompose recognition into preprocessing, text detection, orientation correction, line or word recognition, language correction, and post-processing. Modern deep learning improved these components through document restoration, arbitrary-shaped text detection, sequence recognition, and end-to-end text spotting [10,18,84,88,92,44,61,46,33,43,27]. Recent work further extends OCR toward broader text-centric visual modeling [41,82].

Document intelligence extends OCR beyond character recognition. A document image is a visual, spatial, linguistic, and semantic object. The LayoutLM series [79,78,28] model text tokens with bounding boxes, while later layout-aware models jointly use layout tokens, visual features, and language features for document classification, key information extraction, form understanding, and visual question answering (VQA) [4,70,49,69,38,47,45]. These models often assume text has already been recognized, but they establish an important principle: document understanding requires the interaction of text, layout, and vision.

OCR-free document models then challenged the need for a separate OCR engine. Donut treats visual document understanding as image-to-text generation [31], Pix2Struct uses screenshot parsing pretraining for visual language

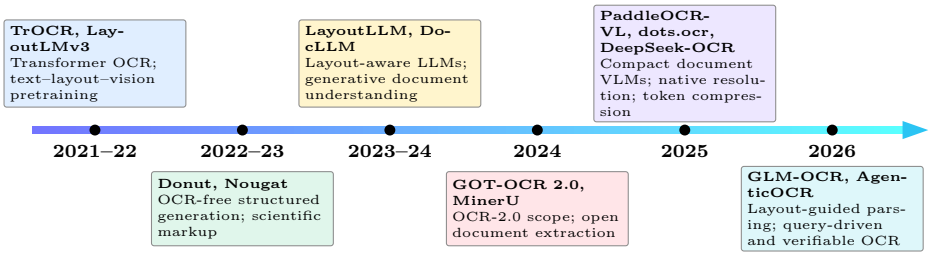

Fig. 1. Evolution of representative OCR large models and document-intelligence systems. The figure highlights dominant technical shifts rather than exhaustively listing all models.

understanding [32], Nougat converts scientific PDFs into markup [6], and DocPedia extends document understanding with large multimodal modeling in the frequency domain [17]. These models are direct ancestors of recent OCR large models because they frame document parsing as structured generation rather than recognition plus rules.

Recent surveys in document intelligence [30] and document parsing [86] highlight the convergence of layout analysis, OCR, VLMs, and LLMs. The literature now increasingly treats document parsing as a systems problem: detect regions, recognize text, infer structure, recover reading order, validate outputs, and connect results to downstream LLM applications. Figure 1 summarizes this trajectory from transformer-based OCR and layout-aware document modeling to OCR-free generation, OCR-specialized VLMs, and grounded or agentic document systems.

2.2 MLLMs as Document Readers

The emergence of general MLLMs created a new baseline for OCR and document reading. Early systems such as Flamingo [2], LLaVA [40], GPT-4V [1], Gemini [63], and Qwen-VL [66] demonstrated unified visual perception and language reasoning. More recent models, including GPT-5.5 [56], Claude Opus 4.7 [3], Gemini 3.1 Pro [23], InternVL3.5 [72], and Qwen3.5 [60], further advance high-resolution perception, long-context processing, and visual document understanding.

However, general MLLMs are not always ideal OCR engines. Their outputs may be semantically plausible but not faithful at the character level. High-resolution pages are token-expensive, and document workflows require evidence such as bounding boxes, confidence, and reproducible reading order. In enterprise settings, privacy, latency, batch processing, and deterministic formatting are equally important. These constraints explain why OCR-specialized large models continue to develop.

3 Data and Evaluation

3.1 Data Engines

Recent OCR progress relies on large-scale diverse data, while traditional datasets are often limited by annotation costs [58,9] or homogeneous formats [90,34]. Data engines have therefore become essential. Training data must vary in language, layout, document type, output format, and supervision granularity. PaddleOCR-VL covers OCR, table, formula, chart, and multilingual tasks [11], while PaddleOCR-VL-1.5 expands both data volume and task coverage [12]. HunyuanOCR reports over 200M image-text pairs across diverse scenes and more than 130 languages [64].

Recent data engines also differ in how supervision is produced. dots.ocr introduces a teacher-student multilingual data engine, where a strong VLM creates structure-preserving multilingual seed documents and a student model scales auto-labeling for pretraining [35]. Dolphin reports more than 30M multi-granularity document parsing samples [19], MonkeyOCR introduces MonkeyDoc across multiple document types [37], SmolDocling creates DocTags for page conversion, charts, tables, formulas, and code [53], and PubMed-OCR contributes domain-specific OCR annotations for PMC open-access articles [26].

Several data lessons are now clear. Synthetic data is indispensable but must preserve real layout complexity; multilingual coverage should include different directionality, connected writing, diacritics, vertical layout, and historical forms; document-level supervision must include reading order and element relations; automatic labels from strong models can propagate systematic mistakes; and evaluation data must be separated carefully from web-scale and synthetic training sources.

3.2 Benchmarks

Benchmarking OCR large models is difficult because tasks differ. Table 1 summarizes representative benchmarks across text recognition, PDF/document parsing, key information extraction, and multilingual document understanding.

OCRBench and OCRBench v2 evaluate general OCR, localization, and reasoning over text-rich images [42,21], while Union14M focuses on large-scale scene text recognition with both labeled and unlabeled data [29]. OmniDocBench and olmOCR-Bench shift the target toward PDF/document parsing and PDF-to-text extraction, where structural fidelity, reading order, tables, formulas, and layout-aware conversion become central [57,59].

Multilingual and real-world document benchmarks further expose robustness gaps. XDocParse, MDPBench, and CC-OCR Bench cover multilingual document parsing, photographed pages, non-Latin scripts, multi-scene text reading, document parsing, and key information extraction [35,36,80]. These benchmarks show that performance can drop sharply outside high-resource, clean, Latin-script, or English-centric settings, and that closed models may be more robust under multilingual and photographic conditions.

Table 1. Representative OCR and document-understanding benchmarks.

Benchmark	Scale	Lang.	Focus
OCRBench [42]	1K QA	Zh–En	General OCR and reasoning
OCRBench v2 [21]	10K public + 1.5K private	Zh–En	Text localization and reasoning
Union14M [29]	4M labeled + 10M unlabeled	N/R	Scene text recognition
OmniDocBench v1.5 [57]	1,355 PDF pages	3	PDF parsing
OmniDocBench v1.6 [57]	1,651 PDF pages	5	PDF parsing
olmOCR-Bench [59]	1.4K PDFs; 7,010 tests	N/R	PDF-to-text extraction
XDocParse [35]	N/R	126	Multilingual document parsing
MDPBench [36]	3,400 images	17	Multilingual document parsing
CC-OCR Bench [80]	7,058 images; 39 subsets	10	KIE, Multilingual document parsing

Note. Lang. = language coverage; Zh–En = Chinese–English; N/R = not reported; KIE = key information extraction.

These benchmarks imply that future evaluation should include multiple layers: recognition fidelity, structural correctness, grounding fidelity, multilingual robustness, and downstream task utility. A system that wins on one layer may fail on another.

4 Methods

4.1 Taxonomy

We define an OCR large model as any system that uses large-scale visual, language, or multimodal modeling to read, parse, convert, or reason over text-rich images and documents. This includes OCR-specialized VLMs, layout-guided parsers, compact document VLMs, and agentic pipelines.

To organize recent work, we group OCR large models into three architectural families according to their dominant system design. Individual techniques such as high-resolution encoding, visual token compression, and verifiable reinforcement learning are treated as cross-cutting mechanisms, because they can be adopted across different families. The three families are:

1. **Layout-guided pipelines.** These systems decompose document understanding into explicit stages, typically layout detection, region recognition, reading-order recovery, and structured output formatting. Many recent systems integrate VLMs into layout-guided pipelines rather than remaining purely modular. Representative examples include MinerU [68], PaddleOCR 3.0 [13], PaddleOCR-VL [11], GLM-OCR [16], Dolphin [19], and the two-stage variant of MonkeyOCR [83].
2. **End-to-end OCR VLMs.** These models map images directly to structured text, such as Markdown, LaTeX, JSON, or document-specific markup, in a single model interface. They replace handcrafted pipelines with a joint image-to-structure generation objective. Representative examples include GOT-OCR 2.0 [74], HunyuanOCR, dots.ocr, LightOnOCR [62], Ocean-OCR [8], and DeepSeek-OCR [75].

3. **Agentic OCR systems.** These systems equip OCR models with the ability to call external tools, verify outputs, or selectively parse content conditioned on a query. They emphasize reliability, controllability, and task-driven execution. Representative examples include DianJin-OCR-R1 [7], ViDoRAG [71], Doc-CoB [51], DocVAL [52], and AgenticOCR [73].

The boundaries are not always sharp. A layout-guided pipeline may adopt an end-to-end VLM as its recognition module, and an agentic system may orchestrate both pipelines and end-to-end OCR VLMs. Nevertheless, the taxonomy helps highlight the dominant design philosophy underlying each system. In the following, we present each family and then discuss cross-cutting techniques, especially high-resolution encoding and visual token compression.

4.2 Layout-Guided Pipelines

Multi-stage pipeline design remains important because it offers controllability, auditability, and compatibility with production workflows. MinerU decomposes document processing into preprocessing, layout detection, OCR, formula/table recognition, reading-order recovery, post-processing, and Markdown/JSON conversion. PaddleOCR 3.0 organizes PP-OCRv5, PP-StructureV3, and PP-ChatOCRv4 into an industrial toolkit. These systems illustrate that deployment cost, hardware coverage, latency, privacy, and downstream integration are as important as raw accuracy.

A newer subtype replaces hand-engineered modules with compact VLMs while keeping explicit layout analysis. PaddleOCR-VL uses PP-DocLayout and a 0.9B VLM for local recognition. GLM-OCR adopts a similar two-stage design, and MinerU2.5 [54] separates global layout analysis from local native-resolution recognition. Dolphin uses heterogeneous anchor prompting. while MonkeyOCR v1.5 [83] uses a Structure–Recognition–Relation triplet.

The main advantage of this family is auditability: explicit layout regions allow evidence alignment, parallel processing, local retry, structured validation, and straightforward engineering deployment. Its limitation is cascading error: a missed region, incorrect reading order, or over-segmented block may propagate downwards. Consequently, recent work focuses on improving layout robustness, error recovery, and cross-page structural consistency while keeping the pipeline efficient.

4.3 End-to-End OCR VLMs

End-to-end OCR VLMs formulate document parsing as conditional image-to-structure generation. Their common premise is to replace explicit module boundaries with a unified objective, so that text, layout, formulas, tables, and semantic relations are learned within the same generative space. GOT-OCR 2.0 broadens this idea into OCR-2.0, where optical symbols include not only text but also formulas, charts, tables, chemical notations, and other artificial visual symbols. HunyuanOCR, dots.ocr, LightOnOCR, Ocean-OCR, Qianfan-OCR, and

FireRed-OCR further show how native-resolution encoding, multilingual data synthesis, instruction tuning, and OCR-specific reinforcement learning can be integrated into a single VLM framework. DeepSeek-OCR pushes this paradigm toward visual token efficiency by studying how compressed visual tokens can recover document content.

The advantage of unified modeling is that layout, text, formulas, tables, and language priors can be jointly learned rather than optimized as isolated modules. It also gives a single model a flexible interface for recognition, parsing, information extraction, and visual question answering. The weakness is generative fragility: a model may hallucinate missing text, omit small regions, repeat content, produce invalid markup, or lose exact spatial grounding. For high-stakes document workflows, this motivates research on grounded outputs, validation, and hybrid control.

4.4 Agentic OCR Systems

While layout-guided pipelines and end-to-end OCR VLMs focus on producing document representations, agentic OCR systems make parsing conditional, interactive, and verifiable. They equip OCR models with the ability to call external tools, validate intermediate outputs, or selectively extract only the information required by a downstream task.

DianJin-OCR-R1 interleaves reasoning with expert OCR tools. ViDoRAG uses OCR in iterative retrieval-augmented reasoning. Doc-CoB and DocVAL study grounded document reasoning and validation, while AgenticOCR performs query-driven parsing by extracting only task-relevant regions.

A key enabler in this family is verifiable reinforcement learning. Because OCR outputs can be automatically checked—text by edit distance, boxes by IoU, LaTeX by parsing or rendering, and tables by structural similarity—rewards derived from such checks provide strong training signals. HunyuanOCR, GLM-OCR, and MonkeyOCR v1.5 report OCR-specific reinforcement learning that uses task-level verifiers to reduce hallucination and improve structural fidelity. This direction points toward a future where OCR models are not merely fluent generators but self-correcting, auditable systems.

4.5 Resolution and Token Compression

Beyond the three architectural families, two technical mechanisms have become pervasive across recent OCR large models. The first is adaptive or native-resolution encoding. Dense documents expose a central tension in VLM design: small text, formulas, and table borders require high-resolution perception, yet high-resolution images increase the visual token budget. Models such as PaddleOCR-VL, HunyuanOCR, dots.ocr, and LightOnOCR adopt native-resolution encoders that preserve page fidelity more faithfully than fixed low-resolution inputs.

The second is visual token compression. DeepSeek-OCR and its successor DeepSeek-OCR 2 [76] explicitly study how a small number of compressed visual

Table 2. Technical comparison of representative OCR large models.

Model	Family	Params	Key techniques
PaddleOCR 3.0[13]	Pipeline	<100M*	PP-OCRv5; PP-StructureV3; PP-ChatOCRv4; distillation
PaddleOCR-VL[11]	Pipeline	0.9B	PP-DocLayoutV2; NaViT encoder; ERNIE-4.5; OTSL/LaTeX
PaddleOCR-VL-1.5[12]	Pipeline	0.9B	PP-DocLayoutV3; 0.9B VLM; text spotting; Markdown/JSON
GLM-OCR[16]	Pipeline	0.9B	PP-DocLayout-V3; region recognition; MTP; SFT/RL
MinerU2.5[54]	Pipeline	1.2B	Coarse-to-fine parsing; global layout; local crops
MonkeyOCR[37]	Pipeline	3B	SRR triplet; structure; recognition; relation modeling
MonkeyOCR v1.5[83]	Pipeline	N/R	Two-stage parsing; reading order; local recognition; VC-RL
dots.ocr[35]	E2E	2.9B	1.2B vision encoder; 1.7B decoder; boxes/classes/text/order
HunyuanOCR[64]	E2E	∼1B	Native ViT; MLP adapter; lightweight LLM; RLVR
Qianfan-OCR[15]	E2E	4B	Layout-as-Thought; image-to-Markdown; boxes/types/order
DeepSeek-OCR[75]	E2E	3B MoE	DeepEncoder; MoE decoder; token compression; OCR-2.0 data
DeepSeek-OCR-2[76]	E2E	N/R	DeepEncoder V2; visual causal flow; semantic ordering
FireRed-OCR[77]	E2E	2B	Qwen3-VL-2B; geometry–semantic data; SFT; constrained GRPO

Note. Pipeline = layout-guided pipeline; E2E = end-to-end OCR VLM; MTP = multi-token prediction; SRR = Structure–Recognition–Relation; VC-RL = visual-consistency reinforcement learning; GRPO = Group Relative Policy Optimization; N/R = not reported. *PaddleOCR 3.0 reports component models with fewer than 100M parameters rather than a single total parameter count.

tokens can recover textual context, enabling efficient processing of long or multi-page documents. The trade-off is information allocation: aggressive compression can discard rare symbols or thin structural cues. The research frontier is therefore moving toward auditable token allocation strategies that preserve evidence for visually small yet semantically important regions, regardless of whether the overarching design is layout-guided, end-to-end, or agentic.

4.6 Representative Model Comparison

Table 2 compares representative OCR parsing models from the layout-guided and end-to-end families. Agentic systems are discussed separately because they often orchestrate tools and retrievers rather than report a single OCR backbone.

Table 3 summarizes reported OmniDocBench v1.6 results of OCR-specialized models and several general-purpose closed MLLMs. The OmniDocBench metrics provide a comparable view of general document parsing performance, whereas the robustness scores are reported by different sources and should be interpreted only as supplementary evidence.

Overall, the OmniDocBench v1.6 results suggest that recent OCR large models have become strong integrated document parsers across text, formulas, and tables. OCR-specialized systems such as MinerU2.5-Pro, GLM-OCR, PaddleOCR-VL-1.5, Qianfan-OCR, and FireRed-OCR achieve strong reported

Table 3. Reported document parsing results of representative OCR-specialized models and general-purpose MLLMs on OmniDocBench v1.6.

Model	Params	Overall↑	Text↓	Formula↑	Table↑
MinerU2.5-Pro[67]	1.2B	95.75	0.036	97.45	93.42
GLM-OCR[16]	0.9B	95.22	0.044	97.18	92.83
PaddleOCR-VL-1.5[12]	0.9B	94.93	0.038	96.89	91.67
PaddleOCR-VL[11]	0.9B	94.18	0.040	95.91	90.65
Qianfan-OCR[15]	4B	93.90	0.040	95.08	90.53
FireRed-OCR[77]	2B	93.26	0.037	95.44	88.04
dots.ocr[35]	3B	90.77	0.048	89.95	87.18
DeepSeek-OCR-2[76]	3B	90.25	0.050	91.84	83.89
HunyuanOCR[64]	1B	89.95	0.088	87.68	91.01
MonkeyOCR-pro-3B[83]	3B	88.57	0.074	88.74	84.35
Gemini 3 Flash[24]	N/R	92.62	0.066	95.16	89.29
Gemini 3 Pro[25]	N/R	92.91	0.064	95.99	89.15
InternVL3.5-241B[72]	241B	83.76	0.130	89.95	74.35
GPT-5.2[55]	N/R	86.59	0.114	88.21	82.95
Kimi K2.5[65]	1T	84.53	0.107	83.50	80.76
Qwen3-VL-235B[5]	235B	89.78	0.063	92.55	83.07

Note. Overall, Text, Formula, and Table are taken from the OmniDocBench v1.6. Overall = Overall score; Text = Text Edit Distance; Formula = Formula CDM; Table = Table TEDS. Text Edit Distance is lower-better, while Overall, Formula, and Table are higher-better.

scores, while general-purpose MLLMs such as Gemini 3 Pro, Gemini 3 Flash, Qwen3-VL-235B, and GPT-5.2 also show competitive document-reading ability. However, these numbers should not be interpreted as a strictly fair or definitive ranking. Even under the same benchmark, final scores can be affected by model or API versions, prompt templates, image preprocessing, decoding strategies, output normalization, and the exact evaluation scripts used. For closed or partially released systems, prediction files, prompts, and full evaluation pipelines may also be unavailable, making exact reproduction and controlled pairwise comparison difficult.

Therefore, Table 3 is more useful for identifying broad trends. The small gaps among several leading systems should be interpreted cautiously, especially when comparing OCR-specialized models with general-purpose MLLMs. Different systems optimize for different priorities, including layout-guided controllability, native-resolution perception, structured Markdown generation, table and formula fidelity, multilingual coverage, and visual-token efficiency. From this perspective, the comparison highlights both overall progress and architectural trade-offs, while also showing that future OCR evaluation should report model versions, prompts, rendering settings, prediction files, and evaluation pipelines more transparently.

5 Challenges and Future Directions

5.1 Reliability: Hallucination, Omission, and Grounding

Hallucination is a persistent problem in large vision-language models [39,48], and generative OCR models are no exception. They may invent plausible words, normalize rare symbols incorrectly, repeat content, or omit small regions. Fu-

ture work should attach text and structure to evidence regions, report uncertainty, and refuse ambiguous regions when necessary. Tool-interleaved verification, verifiable reinforcement learning, and active refusal benchmarks such as MathDoc [91] point toward auditable OCR models.

5.2 Structural Parsing: Tables, Formulas, and Reading Order

Complex tables and formulas remain hard because they require geometry, syntax, semantics, and rendering consistency. Tables may contain merged cells, multi-level headers, footnotes, rotated text, and cross-page continuations, while formulas need precise localization and valid LaTeX [22,89]. Reading-order errors in multi-column layouts, captions, and equations also harm RAG chunks and downstream extraction. Future work should model document-level structure across pages, with benchmarks such as MMLongBench-Doc [50] and LongDocURL [14].

5.3 Multilingual and Long-Tail Documents

Although many models report support for over 100 languages, performance remains uneven because English and Chinese dominate public data and benchmarks. Arabic, Indic, Tibetan, vertical Chinese, historical scripts, classical Chinese [81], and handwriting [87,85] involve different glyph shapes, directions, degradation, and linguistic priors. Photographed and non-Latin documents still need script-balanced data, validated synthetic corpora, layout support, and low-resource benchmarks.

5.4 OCR for RAG, Agents, and Data Production

OCR is becoming infrastructure for RAG, agents, and pretraining data production. PDF conversion systems such as olmOCR [59] remain important, but full-page OCR is not always necessary. For long documents, query-driven agents can build coarse indexes and then parse only task-relevant regions at higher resolution. Future systems may combine coarse indexing with query-conditioned high-resolution parsing.

5.5 Deployment, Reproducibility, and Benchmark Governance

OCR workloads are often large-scale and privacy-sensitive: enterprises process millions of pages, mobile apps may require on-device recognition, and archives may contain confidential records. Compact OCR-specialized models show that smaller models can be competitive when architecture and data are specialized. Reproducibility remains uneven because reports may use private data, changing online model versions, or undisclosed prompts. Future systems should report latency and memory, support local deployment, and adopt versioned benchmarks with prompt disclosure.

For practitioners, model choice should consider auditability, privacy, latency, cost, and structured-output needs, not only leaderboard rank.

6 Conclusion

OCR large models have transformed OCR from narrow text recognition into a foundation for structured document intelligence, RAG, data production, document agents, and knowledge-base construction. Strong systems increasingly combine layout control, high-resolution visual encoding, local VLM recognition, token efficiency, structured output constraints, and verifiable rewards. Future advances progress will depend less on scale alone and more on grounding, controllable structure, data quality, token efficiency, and evaluation aligned with real workflows.

Acknowledgments. This research is supported in part by the National Natural Science Foundation of China (Grant No.: 62476093).

AI-Assisted Research Statement This article was completed through an AI-assisted *Vibe Research* workflow [20]. The human authors defined the survey topic, scope, and methodology, and provided most of the relevant references. The AI system supported material organization and initial drafting. The human authors manually checked the core references, guided the taxonomy and comparative framing, and revised the manuscript through multiple rounds. The final manuscript was reviewed and approved by the authors.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Achiam, J., Adler, S., Agarwal, S., et al.: GPT-4 Technical Report. arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., et al.: Flamingo: A Visual Language Model for Few-Shot Learning. NeurIPS **35**, 23716–23736 (2022)
3. Anthropic: Claude Model Documentation (2026), `https://docs.anthropic.com/en/docs/about-claude/models/overview`
4. Appalaraju, S., Jasani, B., Kota, B.U., et al.: Docformer: End-to-end Transformer For Document Understanding. In: ICCV. pp. 993–1003 (2021)
5. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Blecher, L., Cucurull Preixens, G., Scialom, T., et al.: Nougat: Neural Optical Understanding for Academic Documents. In: ICLR. vol. 2024, pp. 37646–37663 (2024)
7. Chen, Q., Zhang, X., Guo, L., et al.: DianJin-OCR-R1: Enhancing OCR Capabilities via a Reasoning-and-Tool Interleaved Vision-Language Model. arXiv:2508.13238 (2025)
8. Chen, S., Guo, X., Li, Y., et al.: Ocean-OCR: Towards General OCR Application via a Vision-Language Model. arXiv:2501.15558 (2025)
9. Cheng, H., Zhang, P., Wu, S., et al.: M6Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis. In: CVPR. pp. 15138–15147 (June 2023)

10. Cho, H., Wang, J., Lee, S.: Text Image Deblurring Using Text-Specific Properties. In: ECCV. pp. 524–537 (2012)
11. Cui, C., Sun, T., Liang, S., et al.: Paddleocr-vl: Boosting Multilingual Document Parsing Via A 0.9 b Ultra-compact Vision-language Model. arXiv:2510.14528 (2025)
12. Cui, C., Sun, T., Liang, S., et al.: PaddleOCR-VL-1.5: Towards a Multi-Task 0.9 B VLM for Robust In-the-Wild Document Parsing. arXiv:2601.21957 (2026)
13. Cui, C., Sun, T., Lin, M., et al.: Paddleocr 3.0 Technical Report. arXiv:2507.05595 (2025)
14. Deng, C., Yuan, J., Bu, P., et al.: LongDocURL: A Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating. In: ACL. pp. 1135–1159 (2025)
15. Dong, D., Zheng, M., Xu, D., et al.: Qianfan-OCR: A Unified End-to-End Model for Document Intelligence. arXiv:2603.13398 (2026)
16. Duan, S., Xue, Y., Wang, W., Su, Z., Liu, H., Yang, S., Gan, G., Wang, G., Wang, Z., Yan, S., et al.: Glm-ocr technical report. arXiv:2603.10910 (2026)
17. Feng, H., Liu, Q., Liu, H., et al.: DocPedia: Unleashing the Power of Large Multimodal Model in the Frequency Domain for Versatile Document Understanding. Sci. China Inf. Sci. **67**(12), 220106 (2024)
18. Feng, H., Liu, S., Deng, J., et al.: Deep Unrestricted Document Image Rectification. IEEE TMM **26**, 6142–6154 (2024)
19. Feng, H., Wei, S., Fei, X., et al.: Dolphin: Document Image Parsing via Heterogeneous Anchor Prompting. In: Findings of ACL. pp. 21919–21936 (2025)
20. Feng, Y., Liu, Y.: A Visionary Look at Vibe Researching. arXiv:2604.00945 (2026)
21. Fu, L., Kuang, Z., Song, J., et al.: OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning. NeurIPS **38** (2026)
22. Gautam, S., Bhandari, A., Harit, G.: TabComp: A Dataset for Visual Table Reading Comprehension. In: Findings of NAACL. pp. 5773–5780 (2025)
23. Google: Gemini 3.1 Pro: A Smarter Model for Complex Tasks (2026), https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/
24. Google DeepMind: Gemini 3 flash model card. https://deepmind.google/models/model-cards/gemini-3-flash/ (Dec 2025), published: December 2025; Accessed: 2026-06-05
25. Google DeepMind: Gemini 3 pro model card. https://deepmind.google/models/model-cards/gemini-3-pro/ (May 2026), model release: November 2025; last updated: May 2026; Accessed: 2026-06-05
26. Heidenreich, H., Getachew, Y., Dinica, O., et al.: PubMed-OCR: PMC Open Access OCR Annotations. arXiv:2601.11425 (2026)
27. Huang, M., Peng, D., Li, H., et al.: Swintextspotter v2: Towards Better Synergy for Scene Text Spotting. IJCV **133**(8), 5281–5301 (2025)
28. Huang, Y., Lv, T., Cui, L., et al.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: ACM MM. pp. 4083–4091 (2022)
29. Jiang, Q., Wang, J., Peng, D., et al.: Revisiting scene text recognition: A data perspective. In: ICCV. pp. 20543–20554 (2023)
30. Ke, W., Zheng, Y., Li, Y., et al.: Large Language Models in Document Intelligence: A Comprehensive Survey, Recent Advances, Challenges, and Future Trends. ACM TOIS **44**(1), 1–64 (2025)
31. Kim, G., Hong, T., Yim, M., et al.: OCR-Free Document Understanding Transformer. In: ECCV. pp. 498–517 (2022)

32. Lee, K., Joshi, M., Turc, I.R., et al.: Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. In: ICML. pp. 18893–18912. PMLR (2023)
33. Li, M., Lv, T., Chen, J., et al.: TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models. AAAI **37**(11), 13094–13102 (Jun 2023)
34. Li, M., Xu, Y., Cui, L., et al.: DocBank: A Benchmark Dataset for Document Layout Analysis. In: COLING. pp. 949–960 (Dec 2020)
35. Li, Y., Yang, G., Liu, H., et al.: dots.ocr: Multilingual Document Layout Parsing in a Single Vision-Language Model. arXiv:2512.02498 (2025)
36. Li, Z., Lin, Z., Liu, Q., et al.: MDPBench: A Benchmark for Multilingual Document Parsing in Real-World Scenarios. arXiv:2603.28130 (2026)
37. Li, Z., Liu, Y., Liu, Q., et al.: MonkeyOCR: Document Parsing with a Structure-Recognition-Relation Triplet Paradigm. arXiv:2506.05218 (2025)
38. Liao, W., Wang, J., Li, H., et al.: DocLayLLM: An Efficient Multi-modal Extension of Large Language Models for Text-rich Document Understanding. In: CVPR. pp. 4038–4049 (June 2025)
39. Liu, H., Xue, W., Chen, Y., et al.: A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253 (2024)
40. Liu, H., Li, C., Wu, Q., et al.: Visual Instruction Tuning. NeurIPS **36**, 34892–34916 (2023)
41. Liu, Y., Huang, M., Yan, H., et al.: VimTS: A Unified Video and Image Text Spotter for Enhancing the Cross-Domain Generalization. IEEE TPAMI **47**(4), 2957–2972 (2025)
42. Liu, Y., Li, Z., Huang, M., et al.: OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. Sci. China Inf. Sci. **67**(12), 220102 (2024)
43. Liu, Y., Shen, C., Jin, L., et al.: ABCNet v2: Adaptive Bezier-Curve Network for Real-Time End-to-End Text Spotting. IEEE TPAMI **44**(11), 8048–8064 (2022)
44. Long, S., Ruan, J., Zhang, W., et al.: TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. In: ECCV (September 2018)
45. Lu, J., Yu, H., Wang, Y., et al.: A Bounding Box is Worth One Token-Interleaving Layout and Text in a Large Language Model for Document Understanding. In: Findings of ACL. pp. 7252–7273 (2025)
46. Luo, C., Jin, L., Sun, Z.: MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition. Pattern Recogn. **90**, 109–118 (2019)
47. Luo, C., Shen, Y., Zhu, Z., et al.: LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. In: CVPR. pp. 15630–15640 (2024)
48. Lyu, G., Liu, Q., Xu, C., et al.: Revealing and Enhancing Core Visual Regions: Harnessing Internal Attention Dynamics for Hallucination Mitigation in LVLMs. arXiv:2602.15556 (2026)
49. Lyu, P., Li, Y., Zhou, H., et al.: StrucTexTv3: An Efficient Vision-Language Model for Text-Rich Image Perception, Comprehension, and Beyond. arXiv:2405.21013 (2024)
50. Ma, Y., Zang, Y., Chen, L., et al.: MMLongBench-Doc: Benchmarking Long-Context Document Understanding with Visualizations. NeurIPS **37**, 95963–96010 (2024)
51. Mo, Y., Shao, Z., Ye, K., et al.: Doc-CoB: Enhancing Multi-Modal Document Understanding with Visual Chain-of-Boxes Reasoning. arXiv:2505.18603 (2025)
52. Mohammadshirazi, A., Neogi, P.P.G., Lim, S.N., et al.: DocVAL: Validated Chain-of-Thought Distillation for Grounded Document VQA. arXiv:2511.22521 (2025)

53. Nassar, A., Omenetti, M., Lysak, M., et al.: SmolDocling: An Ultra-Compact Vision-Language Model for End-to-End Multi-Modal Document Conversion. In: ICCV. pp. 21972–21983 (2025)
54. Niu, J., Liu, Z., Gu, Z., et al.: MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing. In: ACL Industry Track (2025)
55. OpenAI: Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/ (Dec 2025), accessed: 2026-06-05
56. OpenAI: GPT-5.5 System Card (2026), https://openai.com/index/gpt-5-5-system-card/
57. Ouyang, L., Qu, Y., Zhou, H., et al.: OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations. In: CVPR. pp. 24838–24848 (2025)
58. Pfitzmann, B., Auer, C., Dolfi, M., et al.: DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation. In: KDD. p. 3743–3751 (2022)
59. Poznanski, J., Rangapur, A., Borchardt, J., et al.: olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models. In: Championing Open-source DEvelopment in ML Workshop @ ICML25 (2025), https://openreview.net/forum?id=0RoEHuBzc5
60. Qwen Team: Qwen3.5 Model Documentation and Model Cards (2026), https://qwenlm.github.io/
61. Shi, B., Bai, X., Yao, C.: An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE TPAMI **39**(11), 2298–2304 (2017)
62. Taghadouini, S., Cavaillès, A., Aubertin, B.: LightOnOCR: A 1B End-to-End Multilingual Vision-Language Model for State-of-the-Art OCR. arXiv:2601.14251 (2026)
63. Team, G., Anil, R., Borgeaud, S., et al.: Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 (2023)
64. Team, H.V., Lyu, P., Wan, X., et al.: Hunyuanocr Technical Report. arXiv:2511.19575 (2025)
65. Team, K.: Kimi k2.5: Visual agentic intelligence (2026), https://arxiv.org/abs/2602.02276
66. Team, Q.: Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966 (2023)
67. Wang, B., He, T., Ouyang, L., et al.: MinerU2. 5-Pro: Pushing the Limits of Data-Centric Document Parsing at Scale. arXiv:2604.04771 (2026)
68. Wang, B., Xu, C., Zhao, X., et al.: MinerU: An Open-Source Solution for Precise Document Content Extraction. arXiv:2409.18839 (2024)
69. Wang, D., Raman, N., Sibue, M., et al.: DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. In: ACL. pp. 8529–8548 (2024)
70. Wang, J., Jin, L., Ding, K.: Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In: ACL. pp. 7747–7757 (2022)
71. Wang, Q., Ding, R., Chen, Z., et al.: VidoRAG: Visual Document Retrieval-Augmented Generation via Dynamic Iterative Reasoning Agents. In: EMNLP. pp. 9124–9145 (2025)
72. Wang, W., Gao, Z., Gu, L., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv:2508.18265 (2025)
73. Wang, Z., Ma, D., Zhong, H., et al.: AgenticOCR: Parsing Only What You Need for Efficient Retrieval-Augmented Generation. arXiv:2602.24134 (2026)

74. Wei, H., Liu, C., Chen, J., et al.: General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model. arXiv:2409.01704 (2024)
75. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts Optical Compression. arXiv:2510.18234 (2025)
76. Wei, H., Sun, Y., Li, Y.: DeepSeek-OCR 2: Visual Causal Flow. arXiv:2601.20552 (2026)
77. Wu, H., Lou, H., Li, X., et al.: FireRed-OCR Technical Report. arXiv:2603.01840 (2026)
78. Xu, Y., Xu, Y., Lv, T., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In: ACL-IJCNLP. pp. 2579–2591 (2021)
79. Xu, Y., Li, M., Cui, L., et al.: Layoutlm: Pre-training of text and layout for document image understanding. In: KDD. pp. 1192–1200 (2020)
80. Yang, Z., Tang, J., Li, Z., et al.: CC-OCR: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In: ICCV. pp. 21744–21754 (2025)
81. Yao, X., Wang, M., Chen, B., et al.: WenyanGPT: A Large Language Model for Classical Chinese Tasks. In: IJCAI. pp. 8339–8347 (2025)
82. Yin, L., Xie, X., Li, Z., et al.: MSTAR: Box-free Multi-query Scene Text Retrieval with Attention Recycling. In: NeurIPS. vol. 38, pp. 57856–57881 (2025)
83. Zhang, J., Liu, Y., Wu, Z., et al.: MonkeyOCR v1. 5 Technical Report: Unlocking Robust Document Parsing for Complex Patterns. arXiv:2511.10390 (2025)
84. Zhang, J., Peng, D., Liu, C., et al.: DocRes: A Generalist Model Toward Unifying Document Image Restoration Tasks. In: CVPR. pp. 15654–15664 (June 2024)
85. Zhang, P., Jiang, J., Liu, Y., et al.: MSDS: A Large-Scale Chinese Signature and Token Digit String Dataset for Handwriting Verification. In: NeurIPS. vol. 35, pp. 36507–36519 (2022)
86. Zhang, Q., Wang, B., Huang, V.S.J., et al.: Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. arXiv:2410.21169 (2024)
87. Zhang, Y., Shi, Y., Zhang, P., et al.: MegaHan97K: A Large-Scale Dataset for Mega-Category Chinese Character Recognition with over 97K Categories. Pattern Recognition **167**, 111757 (2025)
88. Zhao, F., Zeng, W., Li, Z., et al.: Uni-DocDiff: A Unified Document Restoration Model Based on Diffusion. In: ACM MM. p. 8204–8213 (2025)
89. Zhao, W., Feng, H., Liu, Q., et al.: TabPedia: Towards Comprehensive Visual Table Understanding with Concept Synergy. NeurIPS **37**, 7185–7212 (2024)
90. Zhong, X., Tang, J., Jimeno Yepes, A.: PubLayNet: Largest Dataset Ever for Document Layout Analysis. In: ICDAR. pp. 1015–1022 (2019)
91. Zhou, C., Tuo, J., Qin, S., et al.: MathDoc: Benchmarking Structured Extraction and Active Refusal on Noisy Mathematics Exam Papers. arXiv:2601.10104 (2026)
92. Zhou, X., Yao, C., Wen, H., et al.: EAST: An Efficient and Accurate Scene Text Detector. In: CVPR (July 2017)