A Review of Applications and Research Advances of Large Models in OCR Technology

Haoye Liu

School of Henan University, Kaifeng, China

Abstract. Optical Character Recognition (OCR) is a key technology for document digitization and textual information extraction. However, traditional methods are often limited by practical challenges such as font diversity, complex layouts, and image degradation. In response, this paper provides a systematic review of recent advances in OCR driven by large models. First, a review framework of "bottleneck analysis – technical response – performance evaluation" is constructed, and the paradigm shift is examined from three dimensions: model architecture, training paradigms, and application scenarios. Within this framework, we systematically categorize and compare three mainstream technical approaches: cross-modal pre-trained models, end-to-end sequence generation methods, and few-shot/zero-shot adaptation techniques. Then, through a comparative analysis of representative models in terms of accuracy, robustness, and efficiency, we highlight the significant advantages of large models in complex document understanding, natural scene text recognition, and multilingual generalization, while also noting their persistent limitations in computational cost, hallucination suppression, and domain-specific adaptation. Finally, by considering emerging trends like lightweight deployment, multimodal fusion, and privacy-preserving computation, the paper envisions OCR evolving toward an efficient, adaptive, and cognitively-enabled new paradigm, thereby providing a roadmap for future research and practice.

Keywords: OCR.Large Language Models. Natural Scene Text Recognition. Systematic Review. Paradigm Shift.

1. Introduction

Optical Character Recognition (OCR) is a pivotal technology for transforming visual text into machine-readable formats, serving as a cornerstone in document digitization [1], finance [2], and healthcare [3]. Despite its evolution, traditional methods relying on template matching and handcrafted features face critical limitations in complex scenarios, including poor generalization against noise and a lack of semantic context for inferring degraded characters [4]. Consequently, the paradigm has shifted toward pre-trained large models (e.g., BERT, GPT) and multimodal architectures (e.g., GPT-4), which integrate bidirectional context to resolve recognition errors. Notably, advanced end-to-end frameworks like VISTA-OCR unify transcription with spatial generation, reducing reliance on intermediate steps and significantly enhancing system robustness and efficiency [6] [7].

1.1. The primary innovations and contributions of this work are as follows

- This study proposes a closed-loop review framework that follows the "problem identification – technical response – performance quantification" sequence, thereby systematically tracing the intrinsic pathway of the paradigm shift induced by large models in OCR.
- This work systematically analyzes the architectural evolution from cross-modal fusion to end-to-end generation. By deconstructing representative models such as LayoutLM and Donut, it reveals the underlying logic of their architectural progression and corresponding performance boundaries.
- This research proposes a "technology-scenario-challenge" framework to analyze the interplay among large-model OCR capabilities, application requirements, and deployment barriers. It further offers practical solutions tailored to resource and privacy constraints.

2. Fundamental Principles and Traditional Methods of OCR

2.1. Technical Workflow of OCR

The OCR technical pipeline generally encompasses five key stages: image preprocessing (such as denoising and binarization), text detection, text recognition, structural parsing, and semantic understanding. Specifically, image preprocessing optimizes input quality, laying the groundwork for text detection, which localizes text regions. Subsequently, text recognition converts visual content into editable characters, while structural parsing categorizes document components like titles, paragraphs, and tables. Finally, semantic understanding extracts critical information and interprets logical correlations. These steps are synergized to transform raw image-based text into data that is readable, editable, and semantically comprehensible in Fig. 1.

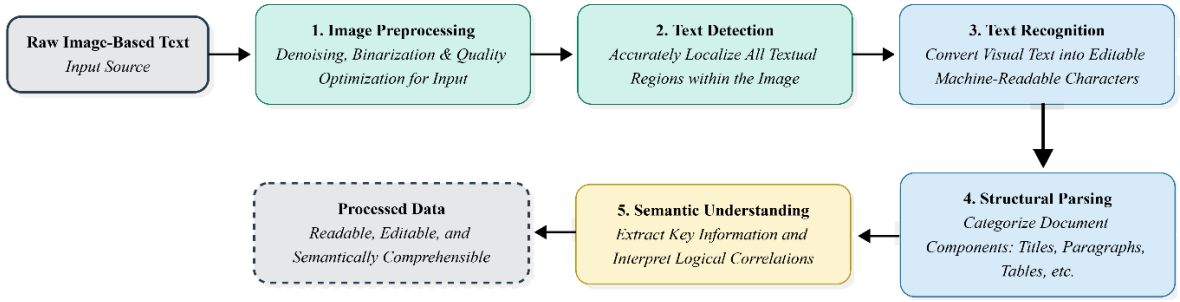


Fig. 1 Flowchart illustrating the stages of OCR technology.

2.1.1. Definition of Large Model OCR

Large Model OCR emerges from the synergy between Optical Character Recognition (OCR) and Large Language Models (LLMs). Fundamentally, it utilizes deep learning models—most notably the Transformer architecture—to realize the end-to-end understanding and processing of visual text. Unlike traditional approaches, this technology adopts a unified framework that consolidates distinct pipeline stages—ranging from image preprocessing, text detection, and recognition to structural parsing and semantic understanding—into a cohesive whole. This integration aims to mitigate the deficiencies of traditional OCR methods, such as cascading errors between modules, limited generalization, and inadequate robustness in complex scenarios.

2.1.2. Technical Essence of Large Model OCR

At its core, Large Model OCR integrates large-scale pre-trained models—notably Large Vision Models (such as Vision Transformer, ViT) and Large Language Models (such as BERT)—to facilitate a vision-language cross-modal learning process. This integration enables a unified, end-to-end "detection-recognition-understanding" workflow, effectively circumventing the dependency on handcrafted features and stepwise tuning required by traditional approaches[5]. Such a paradigm shift markedly improves both performance and robustness, particularly in handling complex environments, diverse typographies, and noise interference[11].

2.2. Review of Traditional Methods

2.2.1. Template Matching

Template matching relies on directly comparing an image of a character against a set of pre-established templates to ascertain its category, typically through similarity calculations. While effective for printed characters, this approach struggles with robustness when confronted by character deformation, noise, or variations in font size.

2.2.2. Handcrafted Feature-based Methods

HOG (Histogram of Oriented Gradients) characterizes edge and shape information by computing histograms of gradient orientations within local regions of an image. This method is capable of

effectively capturing the directionality of character strokes, exhibiting robustness to illumination variations and geometric deformations. Consequently, HOG features are frequently utilized in tasks like pedestrian detection and vehicle wheel hub recognition.

As a local feature descriptor, SIFT(Scale-Invariant Feature Transform) maintains invariance to variations in scale, rotation, and illumination. By identifying keypoints and characterizing the local texture information in their vicinity, SIFT significantly bolsters recognition accuracy. It is especially adept at handling characters subject to varying dimensions or rotational distortions, thereby demonstrating superior performance in image matching and object recognition tasks.

2.2.3. Early Machine Learning Algorithms

SVM classifies samples by identifying the optimal hyperplane in high-dimensional space. In OCR, it often combines with handcrafted descriptors like HOG or SIFT for character classification. For instance, HOG-SVM and SIFT-SVM achieved 0.660 accuracy in rice classification, while HOG-SVM is also widely applied in vehicle, face, and pedestrian detection.

2.3. Overview of Large Model Technologies

2.3.1. Transformer Architecture

As the foundation of modern OCR, the Transformer replaces traditional networks with Self-Attention mechanisms to enable parallel processing and capture long-range dependencies. It utilizes Multi-Head Attention for diverse feature extraction and positional encodings to maintain sequential order, collectively facilitating the robust comprehension of complex visual text.

2.3.2. Pre-training

Pre-training serves as a fundamental phase in Large Model OCR, leveraging massive unlabeled data to establish robust representations via self-supervised tasks like masked prediction. By assimilating visual features, typography, and linguistic contexts, this process captures cross-domain patterns, thereby providing an efficient initialization for downstream fine-tuning that significantly reduces the computational overhead and convergence challenges of training from scratch.

2.3.3. Fine-tuning

Fine-tuning adapts pre-trained models to specific OCR tasks (e.g., handwriting, scene text) using limited labeled datasets. By transferring universal visual-linguistic knowledge and calibrating parameters through supervised learning, this process significantly enhances accuracy, robustness, and efficiency in specialized domains.

3. Key Technologies in Large Model-based OCR

3.1. Cross-modal Fusion Architecture

By facilitating deep interaction between visual and linguistic features, cross-modal fusion architecture significantly boosts the recognition accuracy and comprehension capabilities of Large Model OCR. The underlying technical rationale is twofold:

Firstly, in terms of feature extraction, the system employs the HVFA module and ViT's self-attention mechanism to handle multi-scale visual information, effectively tackling appearance-related challenges like variable font sizes[5]. Concurrently, it harnesses language models refined by reinforcement learning to capture contextual semantics, thereby mitigating noise interference in complex scenarios through mechanisms such as error correction and text augmentation[12].

Secondly, regarding core mechanisms, the architecture hinges on attention mechanisms and feature alignment. It utilizes cross-attention to achieve a profound aggregation of visual and linguistic modalities [13], while simultaneously deploying contrastive learning (e.g., CLIP) to project multimodal representations into a shared embedding space[14]. This ensures precise alignment between specific image regions and their corresponding text sequences.

Microsoft's LayoutLM series (v1/v2/v3) advances document understanding by fusing text semantics, visual features, and layout information[5]. Its multimodal Transformer encoder integrates these attributes into text sequences, leveraging attention mechanisms to capture intricate element interactions[8].

Specifically, LayoutLMv2 introduced a two-stream architecture, whereas LayoutLMv3 implemented a unified Masking strategy. Leveraging pre-training tasks like MVLM for multimodal alignment, these models enhance the parsing of complex structures (e.g., table 1, forms), significantly improving performance on downstream tasks[9] in Fig. 2.

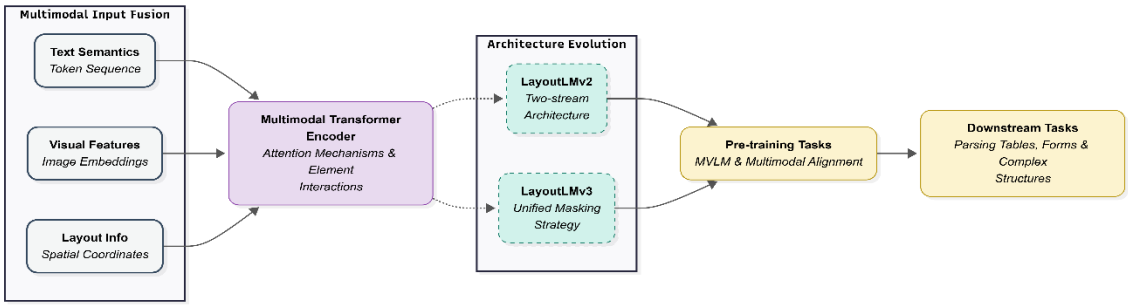


Fig. 2 LayoutLM flowchart.

Donut (Document Understanding Transformer), introduced by NAVER, represents an end-to-end, OCR-free paradigm built entirely on the Transformer architecture. By formulating document understanding as a sequence-to-sequence generation task [10], the model employs an encoder to process raw images and a decoder to directly yield structured text outputs (e.g., JSON). This design bypasses intermediate recognition steps, allowing the model to implicitly learn text detection, transcription, and structural analysis. Furthermore, extensive pre-training on large-scale synthetic data significantly enhances its generalization capabilities across diverse document layouts in Fig. 3.

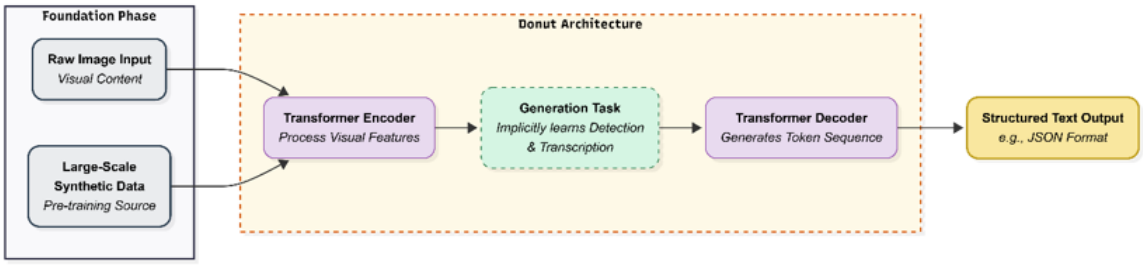


Fig. 3 Donut(NAVER) flowchart.

3.2. Unified End-to-End Models

Unified end-to-end models reframe OCR as a sequence generation problem, replacing traditional multi-stage pipelines to simplify deployment [5]. Architectures like VISTA-OCR leverage Transformer decoders to jointly generate text and coordinates. By eliminating error accumulation inherent in independent components, this strategy enhances robustness and accuracy through global visual-textual modeling.

As a fully Transformer-based architecture, TrOCR replaces CNN backbones and detection stages with a ViT encoder and language model decoder for direct text generation. Supported by large-scale synthetic pre-training, it achieves superior robustness on handwritten and irregular text.

PaddleOCRv3 achieves an industrial-grade balance of accuracy, speed, and size by employing the SVTR algorithm, which merges local CNN features with global Transformer context. Supported by Collaborative Mutual Learning (CML) knowledge distillation, it optimizes parameter efficiency and accuracy. Furthermore, its multi-language support and mobile optimization facilitate robust real-world deployment.

mPLUG-DocOwl advances OCR from perception to cognitive understanding by aligning high-resolution visual features with LLMs via the H-Reducer. Enhanced by instruction tuning, it gains

reasoning and dialogue capabilities to process complex layouts (e.g., nested table 2) and generate structured outputs (e.g., JSON), effectively addressing higher-order analysis tasks.

3.3. Few-Shot and Zero-Shot Adaptation Technologies

Leveraging pre-trained visual-linguistic knowledge, large models adapt to data-scarce scenarios via prompt and transfer learning. This enables few-shot adaptation with minimal samples (10–50) through fine-tuning, and zero-shot recognition via semantic association[16].

Practically, prompt learning employs textual cues, while transfer learning (e.g., CLIP-ViT) utilizes cross-modal similarity for efficient feature propagation[14]. In multilingual settings, models like mBERT address low-resource challenges, and when integrated with visual-layout modalities, facilitate unified cross-modal retrieval frameworks[15].

3.4. Lightweighting and Efficiency Optimization Technologies

The deployment of Large Model OCR on mobile or edge devices is hindered by the massive parameter counts and high computational costs of large-scale pre-trained models. Consequently, achieving an optimal trade-off between accuracy and efficiency necessitates the use of model compression and hardware-software co-optimization strategies.

Techniques such as model distillation, quantization, and pruning are employed to minimize model size and latency, while hardware-software co-design (e.g., FPGAs) further optimizes computation. For instance, LMRPA integrates LLMs into Robotic Process Automation (RPA) workflows, ensuring high performance even in resource-constrained environments[17].

4. Performance Analysis and Applications

4.1. Experiments and Performance Analysis

4.1.1. Performance Comparison on Public Benchmarks

Large Multimodal Models (e.g., GPT-4V, Gemini) excel in tasks like VQA and KIE by leveraging robust visual-linguistic comprehension[18]. However, to mitigate the "hallucination" issue common in generative models, DianJin-OCR-R1 introduces a reasoning and tool-interleaved framework to ensure fidelity[19].

Table 1. A Survey of Key Approaches in Visual Document Understanding.

Model Name	Core Technology	Advantages / Features
GPT-4V/Gemini[5]	Large Multimodal Models (MLLMs)	Powerful visual and language understanding capabilities
DianJin-OCR-R1	Reasoning and tool-interleaved LVLM	Mitigates hallucination issues in generative models
Baseer	Vision-Language Model fine-tuned for Arabic documents	Specially optimized for Arabic OCR; handles complex scripts and layouts
VISTA-OCR	Generative model for unified text detection and recognition	End-to-end efficiency; simultaneously generates text transcriptions and spatial coordinates
HVFA	Multi-scale visual feature aggregation with cross-attention mechanism	Reduces computational costs for MLLMs on multi-scale inputs; enhances robustness
LayoutLM	Multimodal pre-training integrating text, visual, and layout information	Deep understanding of document structure and semantics
Donut (NAVER)	End-to-end OCR-free Transformer	Generates structured text directly from images; eliminates intermediate OCR steps
RL-optimized LLM	Optimization via Reinforcement Learning algorithms	Addresses text diversity and noise interference in complex OCR scenarios

Addressing linguistic complexities (e.g., cursive scripts, diacritics) in Arabic, the Baseer model fine-tunes a VLM on hybrid datasets to enhance performance. In terms of architecture, VISTA-OCR unifies detection and recognition via a Transformer decoder that sequentially generates text and spatial coordinates, improving efficiency by eliminating independent processing branches. Additionally, the HVFA module utilizes cross-attention to aggregate feature pyramids, effectively reducing MLLM computational overhead while robustly handling multi-scale text variations.

4.1.2. Trade-off Analysis: Parameter Size, Inference Speed, and Accuracy (Precision/Recall)

Performance evaluation necessitates balancing parameter efficiency, inference latency, and accuracy[18]. While large-scale MLLMs incur high resource demands, architectural innovations like HVFA and lightweighting techniques (e.g., distillation) effectively optimize deployment. Regarding accuracy, cross-modal architectures (e.g., LayoutLM) and OCR-free end-to-end models (e.g., Donut) excel by fusing features or eliminating error propagation. However, generative approaches introduce hallucination risks, necessitating mitigation strategies such as the interleaved reasoning and tool usage employed by DianJin-OCR-R1[19].

4.2. Analysis of Model Suitability

Table 2. Application Scenarios of Key OCR and MLLM Models

Model Name	Applicable Scenarios
GPT-4V/Gemini	Diverse text-related visual tasks, including text recognition, Visual Question Answering (VQA), and Key Information Extraction (KIE).
DianJin-OCR-R1	OCR tasks necessitating high accuracy and the mitigation of generative errors (e.g., hallucinations).
Baseer	Digitization and information extraction specifically for Arabic documents.
VISTA-OCR	Scenarios requiring simplified OCR pipelines to enhance the efficiency and accuracy of end-to-end processing.
HVFA	Processing document images characterized by varying font sizes and complex layouts.
LayoutLM	Analysis of structured documents, such as form understanding, receipt information extraction, and document classification.
Donut (NAVER)	Document automation, data entry, and Document Visual Question Answering (DocVQA) systems.

5. Future Directions and Challenges

5.1. Future Perspectives

Efficiency: Models will prioritize lightweight, modular designs for edge deployment. Unified architecture (e.g., VISTA-OCR) and optimized aggregation (e.g., HVFA) will be pivotal in reducing computational overhead.

Generalization: Research will target low-resource languages (e.g., Vietnamese, Arabic) via weakly supervised learning and VLM fine-tuning to overcome data scarcity.

LLM Integration: Synergies with generative models (e.g., GPT) will elevate OCR from recognition to understanding (e.g., QA, summarization), utilizing reinforcement learning for adaptive error correction [12].

5.2. Challenges

Adaptability: Achieving robust generalization remains difficult, particularly regarding low-resource languages, artistic typography, and complex visual scenarios (e.g., occlusion, uneven illumination).

Privacy: Stricter regulations necessitate decentralized training. While Federated Learning (FL) offers a solution, challenges regarding accuracy degradation, communication overhead, and heterogeneity persist.

Domain Specificity: Models struggle with vertical domains (e.g., ancient manuscripts, medical reports) characterized by specialized layouts and degradation, where robust information extraction remains an open research problem.

6. Conclusion

This paper systematically examines the paradigm shift driven by large models in Optical Character Recognition (OCR), moving beyond the limitations of traditional methods in handling font diversity, layout complexity, image degradation, and semantic understanding. Using a framework that analyzes bottlenecks, identifies technical responses, and evaluates performance, we demonstrate how large models, via cross-modal fusion, end-to-end generation, and few-shot learning, significantly enhance robustness and semantic comprehension in complex scenarios. While challenges remain in computational efficiency, hallucination suppression, and domain-specific adaptation, these models show considerable promise in practical applications like document digitization, financial processing, and medical record analysis. Future research should focus on lightweight deployment, multilingual generalization, integration with generative models, and privacy-preserving frameworks, ultimately advancing OCR from perceptual extraction toward genuine cognitive understanding.

References

[1] A. Jain and J. Sharma, "Classification and interpretation of characters in multi-application OCR system," *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, Delhi, India, 2014, pp. 1-6, https://doi.org/10.1109/ICDMIC.2014.6954231

[2] Xia, S. (2023). Application of Python automation tool combined with OCR technology in "financial checkup" of universities. *Accounting and Corporate Management*, *5*(12), 130-136, https://doi.org/10.23977/acccm.2023.051220

[3] James, J. K., Maran, T., Rice, M. P., Hunt, T. S., Peterson, K. J., Hogan, W. J., Damani, S., & Ryu, A. J. (2024). Experience with an optical character recognition search application for review of outside medical records. *Mayo Clinic Proceedings: Digital Health*, *2*(4), 511-514. https://doi.org/10.1016/j.mcpdig.2024.08.001

[4] Vij, S., Jain, A., Tayal, D., Kumar, V., Arora, R., & Arora, R. (2025). Adaptive OCR error correction for handwritten texts: A semantic and statistical approach. *International Journal of* Information *Technology*, *17*(7), 4371–4377. https://doi.org/10.1007/s41870-025-02627-5

[5] Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.-Y., Wang, Y., Tian, Y., & Gao, W. (2023). Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey. *Machine Intelligence Research*, *20*(4), 447–482. https://doi.org/10.1007/s11633-022-1410-8

[6] Yao, J., Wang, X., Yang, S., & Wang, B. (2024). ViTMatte: Boosting image matting with pre-trained plain vision transformers. Information Fusion, 103, 102091. https://doi.org/10.1016/j.inffus.2023.102091

[7] Wang, Y., He, J., Wang, D., Wang, Q., Wan, B., & Luo, X. (2024). Multimodal transformers with adaptive modality weighting for multimodal sentiment analysis. Neurocomputing, 572, 127181. https://doi.org/10.1016/j.neucom.2023.127181

[8] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2019). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. arXiv. https://doi.org/10.48550/ARXIV.1912.13318

[9] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., & Zhou, L. (2020). LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2012.14740

[10] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2021). OCR-free Document Understanding Transformer (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2111.15664

[11] Xue, Y. (2023). Simulation Research on Large Language Model of Complex OCR Scene Based on Reinforcement Learning Algorithm Optimization. 2023 International Conference on Internet of Things,

Robotics and Distributed Computing (ICIRDC), 738–742. https://doi.org/10.1109/icirdc62824.2023.00140

[12] Xue, Y. (2023). Simulation Research on Large Language Model of Complex OCR Scene Based on Reinforcement Learning Algorithm Optimization. 2023 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC), 738–742. https://doi.org/10.1109/icirdc62824.2023.00140

[13] Lin, J., Yang, A., Zhang, Y., Liu, J., Zhou, J., & Yang, H. (2020). InterBERT: Vision-and-Language Interaction for Multi-modal Pretraining (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2003.13198

[14] Kim, B., Choi, S., Hwang, D., Lee, M., & Lee, H. (2023). Transferring Pre-trained Multimodal Representations with Cross-modal Similarity Matching (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2301.02903

[15] Yi, W., Cai, X., Ma, H., Fu, Z., & Zhan, Y. (2025). A Library-Oriented Large Language Model Approach to Cross-Lingual and Cross-Modal Document Retrieval. Electronics, 14(15), 3145. https://doi.org/10.3390/electronics14153145

[16] Chen, Y.-C., Li, W.-H., & Chen, C.-S. (2024). Open-Vocabulary Panoptic Segmentation Using Bert Pre-Training of Vision-Language Multiway Transformer Model. 2024 IEEE International Conference on Image Processing (ICIP), 2494–2500. https://doi.org/10.1109/icip51287.2024.10647459.

[17] Abdellaif, O. H., Hassan, A. N., & Hamdi, A. (2025). *LMRPA: Large Language Model-Driven Efficient Robotic Process* Automation *for OCR*. *IEEE Transactions on Automation Science and Engineering*. Advance online publication. https://arxiv.org/abs/2412.18063.

[18] Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.-C., Liu, C.-L., Jin, L., & Bai, X. (2024). OCRBench: on the hidden mystery of OCR in large multimodal models. Science China Information Sciences, 67(12). https://doi.org/10.1007/s11432-024-4235-6.

[19] Chen, Q., Zhang, X., Guo, L., Chen, F., & Zhang, C. (2025). DianJin-OCR-R1: Enhancing OCR Capabilities via a Reasoning-and-Tool Interleaved Vision-Language Model (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2508.13238.