QARI-OCR: High-Fidelity Arabic Text Recognition through Multimodal Large Language Model Adaptation

Anonymous ACL submission

Abstract

The inherent complexities of Arabic script—its cursive nature, diacritical marks (*tashkīl*), and varied typography—pose persistent challenges for Optical Character Recognition (OCR). We present Qari-OCR, a series of vision-language models derived from Qwen2-VL-2B-Instruct, progressively optimized for Arabic through iterative fine-tuning on specialized synthetic datasets. Our leading model, QARI v0.2 achieves a the strongest performance with a Word Error Rate (WER) of 0.160, Character Error Rate (CER) of 0.061, and BLEU score of 0.737 on diacritically-rich texts. Qari-OCR demonstrates the strongest handling of *tashkīl*, diverse fonts, and document layouts, alongside impressive performance on low-resolution images. Further explorations (QARI v0.3) showcase strong potential for structural document understanding and handwritten text. This work delivers a marked improvement in Arabic OCR accuracy and efficiency, with all models and datasets released to foster further research.

1 Introduction

Digital text accessibility is fundamental to information preservation, dissemination, and large-scale analysis in today's data-driven society. Optical Character Recognition (OCR) has achieved remarkable success for Latin-based scripts; however, complex writing systems such as Arabic continue to pose substantial challenges. Arabic script is inherently cursive, exhibits context-dependent character shapes, employs a rich system of diacritical marks (*tashkīl*), and spans a wide range of typographic styles. These properties collectively complicate character segmentation, visual discrimination, and sequence modeling, limiting the effectiveness of conventional OCR pipelines (Al-Sheikh et al., 2020).

The difficulty of Arabic OCR is not merely a linguistic issue, but also a methodological one that intersects with recent developments in vision-language modeling. Advances in multimodal large language models (MLLMs) invite a reexamination of the role played by visual encoders within language-centric systems (Goyal et al., 2017; Kazemzadeh et al., 2014). Rather than viewing vision modules as generic perceptual front-ends for tasks such as visual question answering, we adopt an LLM-centric perspective in which visual representations are optimized to support efficient, faithful textual reasoning. From this viewpoint, OCR constitutes a particularly well-defined intermediate modality: it instantiates a natural compression-decompression process, mapping dense visual inputs into structured linguistic representations that large language models can directly consume. Crucially, OCR also offers clear semantic objectives and standardized quantitative evaluation metrics, making it an effective testbed for studying vision-language interaction under controlled yet realistic conditions.

Arabic provides a uniquely demanding setting for this paradigm. Spoken by over 420 million people worldwide, Arabic plays a central role in cultural preservation, religious scholarship, and historical documentation (UNESCO, 2024). Yet existing Arabic OCR systems consistently underperform relative to their Latin-script counterparts, with especially poor handling of diacritics—elements that are essential for correct pronunciation, grammatical interpretation, and semantic disambiguation (Alwajih et al., 2024). These shortcomings are amplified in real-world documents that feature heterogeneous fonts, dense layouts, degraded scans, or classical and fully vocalized text.

In this work, we introduce *Qari-OCR*, a family of vision–language OCR models specialized for high-fidelity Arabic text recognition. Built upon the *Qwen2-VL-2B-Instruct* backbone, Qari-OCR is developed through an iterative fine-tuning strategy using progressively enriched synthetic datasets,

Table 1: Key Characteristics and Objectives of Qari-OCR Model Versions.

Model Ver.	Key Features/Focus	Objective/Tested Capability	Training Dataset Size	HTML?	Diacritics?	Layout Complexity?	Handwritten Support?
Qari-OCR v0.1	Clean, no diacritics, 5 fonts, uniform min. size/layout.	Baseline on legible, low-noise data.	5,000	✗	✗	✗	✗
Qari-OCR v0.2	Diacritics, broader typography (10 fonts), linguistic complexity.	Recognition of diacritic-rich/classical text.	50,000	✗	✔	✗	✗
Qari-OCR v0.3	Multi-font sizes/page (headers, body), realistic layouts.	Spatial parsing for mixed-size, complex layouts.	10,000	✔	✔	✔	✔

each designed to target specific challenges of Arabic script. To support transparency, reproducibility, and future research, we release the complete synthetic data generation pipeline, model training procedures, and evaluation code in an anonymous repository at `https://anonymous.4open.science/r/QARI-OCR-60A2/README.md`.

2 Related Work

Optical Character Recognition (OCR) has evolved from early rule-based pipelines toward end-to-end neural approaches, with each generation addressing limitations exposed by complex scripts such as Arabic. Traditional OCR systems relied on explicit preprocessing and character segmentation, which proved brittle for Arabic due to its cursive structure and context-dependent letter forms (Alrobah and Albahli, 2022).

Deep learning substantially improved OCR performance by enabling implicit segmentation and sequence modeling. CNN–RNN architectures with CTC loss, such as CRNN (Puigcerver, 2017), and later transformer-based models like TrOCR (Li et al., 2023), achieved strong results on general text. However, these models were primarily developed and evaluated on Latin scripts, limiting their effectiveness for Arabic without significant adaptation.

To address this gap, Arabic-specific OCR research incorporated targeted datasets and script-aware modeling choices (Yousef et al., 2020). More recent foundation models, such as Qalam (Bhatia et al., 2024), extend this direction by leveraging multimodal encoders and large-scale Arabic supervision for printed and handwritten text. While effective, these systems are often specialized and computationally demanding.

In parallel, industrial OCR toolkits have emphasized efficiency and deployability. PaddleOCR 3.0 (Cui et al., 2025) integrates multilingual text recognition with document parsing and information extraction using compact models that prioritize inference speed and engineering robustness. Although not designed as multimodal language models, such systems provide strong practical baselines for OCR accuracy under resource constraints.

The latest shift involves Multimodal Large Language Models (MLLMs), which unify vision and language understanding within a single framework. Models such as Qwen2-VL (Wang et al., 2024) and AIN (Heakl et al., 2025) enable OCR as one of many capabilities, but their general-purpose design often limits performance on high-fidelity Arabic text recognition, particularly for diacritics and diverse typography.

Our work, *Qari-OCR*, builds on this landscape by specializing a general-purpose MLLM—*Qwen2-VL-2B-Instruct*—for Arabic OCR through targeted synthetic data generation and parameter-efficient fine-tuning. By jointly addressing diacritic accuracy, font diversity, and document realism, Qari-OCR advances high-fidelity Arabic text recognition while retaining the flexibility of multimodal language models. A comparative summary of prior approaches and their capabilities is provided in Table 2.

3 Methodology

The development of Qari-OCR was implemented through a two-stage methodological framework: firstly, the generation of diverse synthetic datasets

Table 2: Evolution of OCR Approaches and Key Capabilities Relevant to Arabic.

OCR Approach Category	End-to-End	Arabic Diacritics	Font / Style Diversity	Multimodal	Representative Models
Traditional OCR	✗	✗	✗	✗	Tesseract (early versions)
CNN–RNN OCR	✔	Ltd.	Ltd.	✗	CRNN (Puigcerver, 2017)
Transformer-based OCR	✔	Ltd.	Ltd.	✗	TrOCR (Li et al., 2023)
Arabic-Specific OCR	✔	Ltd.	✔	✗	Arabic DL OCR (Yousef et al., 2020)
Arabic Foundation OCR Models	✔	✔	✔	✔	Qalam (Bhatia et al., 2024)
Industrial OCR Systems	✔	Ltd.	✔	✗	PaddleOCR 3.0 (Cui et al., 2025)
General MLLMs	✔	Ltd.	Ltd.	✔	Qwen2-VL (Wang et al., 2024)
Arabic-Inclusive MLLMs	✔	✔	✔	✔	AIN (Heakl et al., 2025)
Qari-OCR (This Work)	✔	✔	✔	✔	**QARI v0.1 / v0.2 / v0.3**

engineered to encapsulate the complexities of Arabic script; and secondly, the iterative fine-tuning of an advanced vision-language model using these specialized datasets. An illustrative overview of this workflow is presented in Figure 1.

3.1 Synthetic Dataset Generation for QARI

To bridge gaps in existing Arabic OCR corpora—namely diacritic coverage, font diversity, and realistic layouts—we devised a three-stage synthetic data pipeline. Two complementary text sources were used: a modern news article collection and a classical Islamic corpus (rich in *tashkīl*). The text was rendered programmatically in HTML using twelve distinct Arabic fonts (from common Naskh to ornate calligraphic styles) at sizes varying between 14 px and 100 px, then converted to PDF via *WeasyPrint* [1] and to images via *pdf2image* [2].

- **Dataset v0.1:** Non-diacritized text, a limited font set, and uniform minimal size establish a high-legibility baseline.
- **Dataset v0.2:** The dataset v0.2 introduces full diacritics and expands the font repertoire to enhance the recognition of vocalized and classical texts.
- **Dataset v0.3:** Introduces mixed font sizes on each page to simulate realistic document structures (headers, body, annotations) and HTML spatial/layout parsing.

Finally, each image undergoes one of three synthetic degradation treatments—*Clean*, *Moderately Degraded* (subtle noise, color shifts, mild blur), or *Heavily Degraded* (textured backgrounds, aggressive blur)—with all variants paired to their ground-truth transcription. This progression yields a robust, multi-faceted Arabic OCR dataset suitable for training and evaluating Qari-OCR across increasing levels of linguistic, typographic, and visual complexity.

3.2 Model Architecture and Training Strategy

We built Qari-OCR on the *Qwen2-VL-2B-Instruct* backbone (Wang et al., 2024), leveraging its Naive Dynamic Resolution for adaptive image scaling and M-RoPE for robust cross-modal positional embeddings. To optimize fine-tuning efficiency, we optionally quantized the model to 4-bit and inserted LoRA adapters (rank = 16) into both vision and language modules.

Training data comprised conversationally formatted image–text pairs, where each "user" message carried an image and prompt, and the "assistant" reply provided the ground-truth Arabic transcription. We conducted three matched fine-tuning runs, each on a different synthetic dataset version, as summarized in 1.

All models were fine-tuned for a single epoch using the Unsloth library.[3] with the AdamW optimizer (Loshchilov and Hutter, 2017) and with $learning\_rate$ equal to 2e-4 and $weight\_decay$

[1] https://weasyprint.org
[2] https://pdf2image.readthedocs.io/en/latest/index.html
[3] https://github.com/unslothai/unsloth

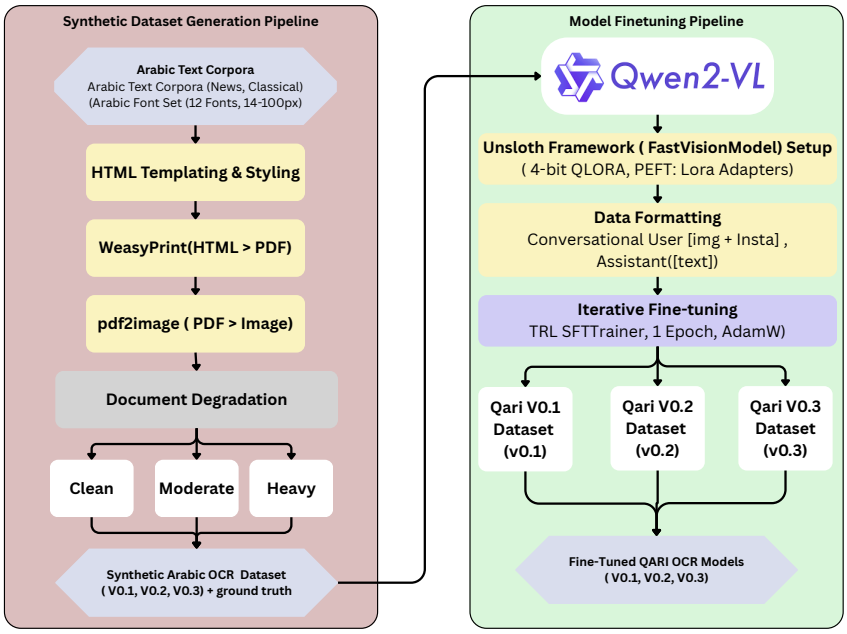

Figure 1: Qari-OCR Dataset Generation and Model Training Pipeline

of 0.01 with linear *lr_scheduler*. Input images were resized and normalized to Qwen2-VL specifications, and training was orchestrated with Hugging Face's`SFTTrainer`[4] using the `UnslothVisionDataCollator`, a per-device batch size of 2, and 4 gradient-accumulation steps (effective batch size = 8). All experiments ran on a single NVIDIA A6000 GPU (48 GB VRAM).

4 Experimental Results

This section describes the experimental setup, evaluation protocol, and empirical results used to benchmark Qari-OCR against representative OCR baselines on challenging Arabic text.

We constructed a test set consisting of 200 scanned pages drawn from traditional Arabic printed materials, encompassing fully vocalized text, complex ligatures, dense line spacing, and heterogeneous layouts. This dataset was designed to reflect the characteristics of historical, religious, and scholarly documents where Arabic OCR systems often struggle. To ensure a fair comparison, all images were processed using the same generic preprocessing pipeline, with no language-specific heuristics, layout annotations, or manual corrections applied, thereby evaluating each system's raw recognition capability.

Our benchmark suite includes Qari-OCR and a diverse set of baseline OCR systems spanning classical engines, industrial solutions, and recent vision–language models. Specifically, we evaluate Tesseract OCR (Smith, 2007), EasyOCR (Pattanayak et al., 2023), Mistral OCR (Mistral AI Team, 2025), AIN (Heakl et al., 2025), Qwen 2.5–7B Instruct, and Qwen 2–7B (Wang et al., 2024). This selection enables a comprehensive comparison across different modeling paradigms, ranging from rule-based and convolutional architectures to large multimodal language models. We include Mistral OCR as a reference point for current commercial OCR performance; however, our primary comparisons focus on open-source models evaluated under identical conditions.

To quantitatively assess OCR performance on Arabic text, we employ three complementary evaluation metrics: Character Error Rate (CER), Word Error Rate (WER) (Klakow and Peters, 2002), and BLEU score (Papineni et al., 2002). CER computes the normalized Levenshtein distance at the character level between predicted and ground-truth transcriptions and is particularly sensitive to diacritic errors and morphologically complex character sequences that are critical in Arabic. WER measures recognition accuracy at the word level, capturing segmentation and substitution errors that affect sentence structure and readability. The BLEU score evaluates n-gram overlap between predictions and references, providing an indication of phrase-level fidelity and overall linguistic coherence. Together, these metrics offer a comprehen-

[4] https://huggingface.co/docs/trl/en/sft_trainer

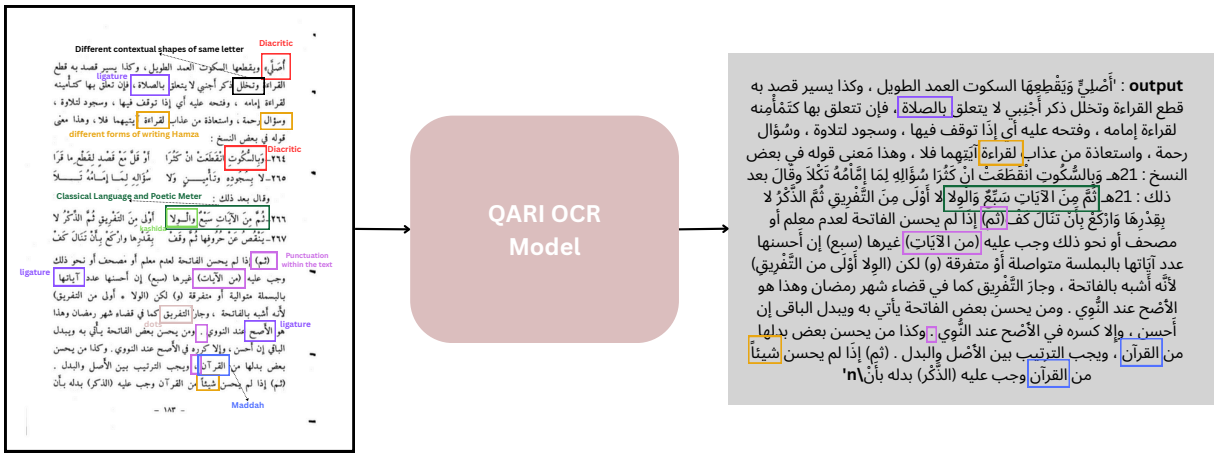

Figure 2: Qualitative example demonstrating Qari-OCR's handling of various Arabic script complexities. The input image (left, with annotations highlighting features like diacritics, ligatures, contextual shapes, etc.) is processed by the Qari-OCR model, producing the transcribed text output (right).

sive assessment of fine-grained accuracy, structural correctness, and semantic preservation in Arabic OCR outputs. While BLEU is traditionally used in machine translation, we employ it here to quantify longer-range linguistic coherence, particularly for evaluating sentence-level coherence beyond isolated word recognition.

Table 3: Comparative performance of OCR models on the Arabic test set. Lower CER/WER and higher BLEU indicate better performance.

Model	CER ↓	WER ↓	BLEU ↑
Tesseract OCR	0.436	0.889	0.108
EasyOCR	0.791	0.918	0.051
AIN	0.640	0.830	0.210
Qwen 2.5-7B Instruct	0.550	0.800	0.220
Qwen 2-7B	0.740	1.050	0.160
PaddleOCR-VL	0.480	0.780	0.248
LightOnOCR-1B-1025	0.990	1.402	0.108
Mistral OCR (API-based)	0.210	0.440	0.570
QARI v0.1 (Ours)	1.915	2.025	0.221
QARI v0.2 (Ours)	**0.061**	**0.160**	**0.737**
QARI v0.3 (Ours)	0.300	0.485	0.545

The comparative performance of our Qari-OCR model variants (QARI v0.1, v0.2, and v0.3) and selected baseline systems was evaluated on the Arabic test set. Quantitative results in terms of Character Error Rate (CER), Word Error Rate (WER), and BLEU score are reported in Table 3.

As shown in Table 3, **QARI v0.2** achieves the strongest performance among evaluated open-source systems, establishing a a strong benchmark with a CER of 0.061, a WER of 0.160, and a BLEU score of 0.737. These gains highlight the effectiveness of our targeted multimodal fine-tuning strategy, particularly the use of synthetic training data enriched with full diacritical coverage and diverse typographic variations (Dataset v0.2). QARI v0.2 also surpasses the API-based Mistral OCR across all three metrics, despite Mistral's strong performance among non-specialized systems.

Among the additional baselines, PaddleOCR-VL demonstrates moderate performance, achieving a BLEU score of 0.248, but remains significantly behind Qari-OCR in both character- and word-level accuracy. LightOnOCR exhibits the weakest results, with a WER of 1.402 and a CER approaching 1.0, indicating extensive insertion and substitution errors and limited robustness on diacritically rich Arabic text. The general-purpose Qwen models, when used without task-specific adaptation, likewise show higher error rates, reinforcing the necessity of specialized training for complex scripts such as Arabic.

Notably, QARI v0.3 achieves a favorable balance between accuracy and structural robustness, attaining a WER of 0.485 and a BLEU score of 0.545. The gap between CER and WER for Qari-OCR models suggests that residual errors are predominantly localized word-level substitutions rather than complete structural failures. Overall, these results demonstrate that Qari-OCR delivers the strongest accuracy, stability, and linguistic fidelity compared to both lightweight industrial OCR systems and general-purpose vision–language models.

Figure 2 provides a visual illustration of Qari-OCR's output on a challenging text sample. The input image (left panel of Figure 2) exhibits several

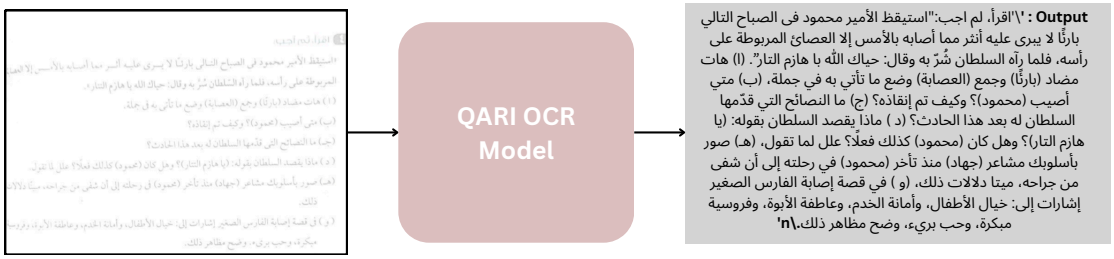

Figure 3: Example of Qari-OCR (v0.3) accurately transcribing Arabic text from a low-resolution and tightly cropped image, showcasing robustness to visual constraints.

Table 4: CER, WER, and BLEU Score results by Font and Model on SARD Dataset

Metric	Model	Amiri	Arial	Calibri	Sakkal M.	Scheherazade
CER↓	Mistral OCR	**0.011**	**0.051**	**0.035**	**0.040**	**0.020**
	Qari v0.2	0.200	0.230	0.193	0.216	0.156
	Qari v0.3	0.350	0.461	0.400	0.424	0.483
WER↓	Mistral OCR	**0.041**	**0.248**	**0.166**	**0.194**	**0.099**
	Qari v0.2	0.267	0.308	0.249	0.293	0.211
	Qari v0.3	0.369	0.482	0.432	0.449	0.464
BLEU↑	Mistral OCR	**0.920**	0.634	0.746	0.715	**0.845**
	Qari v0.2	0.723	**0.703**	**0.745**	**0.701**	0.782
	Qari v0.3	0.346	0.229	0.286	0.279	0.255

Figure 4: Qari-OCR v0.3 successfully transcribing handwritten Arabic text, maintaining sentence structure, punctuation, and recognizing itemized formatting.

features typical of printed Arabic that pose difficulties for OCR systems. These include the full array of diacritics (*tashkīl*) essential for pronunciation and meaning; ligatures such as Lam-Alif (ﻻ); contextually variant letterforms; classical language structures and poetic meter conventions; embedded punctuation and Eastern Arabic numerals; diverse orthographic forms of the Hamza (ء); and features like Maddah (آ) and crucial letter-distinguishing dots.

The corresponding output from our Qari-OCR model (right panel of Figure 2) showcases a high degree of fidelity in transcribing these intricate elements. The model proficiently recognizes the majority of diacritical marks, accurately segments words despite ligatures and contextual letter shaping, and correctly renders classical linguistic forms. This qualitative performance provides strong corroborative evidence for the quantitative results, especially for QARI v0.2, highlighting its robustness in managing the various challenges frequently encountered in real-world Arabic textual scripts.

Beyond quantitative benchmarks, qualitative analysis is crucial for understanding the model's practical capabilities. Figure 2 illustrates Qari-OCR's proficiency in handling different complexities, supporting the strong quantitative performance of QARI v0.2.

Furthermore, the model's resilience to optical degradation and its ability to handle varied inputs were tested. As shown in Figure 3, Qari-OCR (specifically QARI v0.3, trained on more complex layouts) accurately transcribes text from a low-resolution image. Despite the image's small size and tightly cropped boundaries, the model robustly detects and transcribes the Arabic text, demonstrating its effectiveness with compressed layouts, edge-bound scripts, and reduced-resolution content. This capability is vital for digitizing real-world historical or educational Arabic materials, which may not always be of pristine quality.

In addition to printed text, QARI v0.3 was also assessed for its ability to process handwritten Arabic, a notoriously challenging task. Figure 4 il-

lustrates its performance on a handwritten sample. The model accurately detects full sentences, preserving punctuation and word boundaries. Notably, it correctly interprets visual structural cues, such as itemized lists (akin to bullet points) and sentence-level formatting, even with the inherent variability of handwriting. This shows promising initial capabilities for handling handwritten Arabic content.

These qualitative examples, particularly from QARI v0.3 which was trained on more diverse layouts, complement the quantitative results and highlight the practical utility of Qari-OCR in handling a range of challenging real-world Arabic document types.

To evaluate robustness across diverse Arabic fonts, we benchmarked the best-performing models, including QARI v0.2, QARI v0.3, and Mistral OCR, on the SARD dataset[5], which includes 1,000 images spanning five common fonts including; Amiri, Arial, Calibri, Sakkal Majalla, and Scheherazade.

As shown in Table 4, Mistral achieved the lowest error rates overall, particularly excelling in CER and WER. However, QARI v0.2 was highly competitive—outperforming Mistral OCR in BLEU for the Arial font and matching it closely for Calibri. Notably, QARI v0.2's BLEU scores outperformed Mistral OCR for some fonts, including Arial, Calibri, and Sakkak, and consistently outperformed QARI v0.3 across all metrics. These results highlight QARI v0.2 as a strong open-source alternative, balancing accessibility, performance, and versatility across typographic variations.

Moreover, to assess the trade-offs between model size, computational efficiency, and performance, we evaluated different quantization levels for our QARI v0.2 and QARI v0.3 models. Specifically, we compared versions fine-tuned or inferred using 8-bit precision against those utilizing more aggressive 4-bit quantization. The results, presented in Table 5, highlight the impact of these quantization strategies on the CER, WER, and BLEU scores.

As observed in Table 5, employing 8-bit quantization during fine-tuning or inference maintains strong performance for both QARI v0.2 and QARI v0.3, offering a good balance between efficiency and accuracy. However, the more aggressive 4-bit quantization leads to a substantial degradation in performance across all metrics for both model versions. This suggests that while 4-bit quantization significantly reduces the model footprint and can accelerate inference, it incurs a considerable accuracy cost for the fine-grained task of Arabic OCR with these specific models and fine-tuning parameters. The 8-bit versions, therefore, represent the more practical choice when accuracy is paramount, while 4-bit might be considered only in scenarios with extreme computational constraints where a significant drop in accuracy is acceptable.

[5]https://huggingface.co/datasets/riotu-lab/SARD

Table 5: Performance of QARI-OCR with 8-bit Vs. 4-bit Quantization.

Model	Quant.	CER ↓	WER ↓	BLEU ↑
QARI v0.2	8-bit	0.091	0.255	0.583
	4-bit	3.452	4.516	0.001
QARI v0.3	8-bit	0.133	0.353	0.472
	4-bit	3.228	6.428	0.001

5 Discussion

Our experiments reveal distinct strengths across the Qari-OCR model iterations. While QARI v0.2, trained on 50,000 diverse samples (Dataset v0.2), demonstrates the strongest overall quantitative performance for plain text recognition (Table 3), QARI v0.3, developed with a smaller 10,000-sample dataset focused on complex HTML-like layouts (Dataset v0.3), excels in preserving document structure.

Qualitative analysis, as shown in Figure 5, illustrates that QARI v0.3 effectively reconstructs HTML tags and formatting from input images, often achieving lower local error rates on these structurally rich examples compared to QARI v0.2's plain text output. This proficiency stems directly from QARI v0.3's targeted training on layout-aware synthetic data. The trade-off appears to be that QARI v0.2's larger and more varied character-level training data fostered better general textual accuracy, whereas QARI v0.3's smaller, specialized dataset, combined with a single training epoch, prioritized structural fidelity, potentially at the cost of some raw text accuracy on average.

Furthermore, resource efficiency considerations favor the QARI v0.3 approach for structure-oriented tasks. As depicted in Figure 6, the 10k-sample training regimen (QARI v0.3's development) was significantly more economical in terms of training time and estimated CO2 emissions (1.88 kg eq. CO2 over 11 hours) compared to the 50k-

Figure 5: Qualitative comparison of QARI v0.2 and QARI v0.3 outputs against Input and Ground Truth for various Arabic text samples.

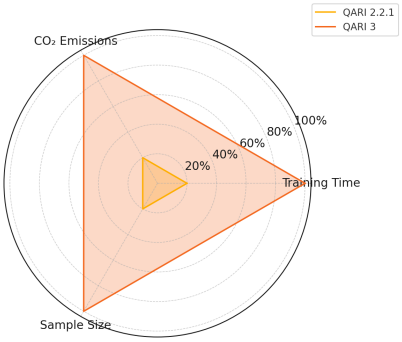

Figure 6: Comparison of estimated resource consumption (CO2 Emissions, Training Time, Sample Size) for training QARI-OCR model variants.

sample training (represented by (represented by QARI v0.2's development, at 9.4 kg eq. CO2 over 55 hours). This highlights the potential for developing specialized, efficient models when the primary objective is structural document conversion.

In essence, QARI v0.2 serves as our most robust general-purpose Arabic OCR engine for accurate plain text extraction. QARI v0.3, however, validates a promising and resource-efficient strategy for applications requiring the understanding and reproduction of document structure, like HTML. The optimal model choice is therefore contingent on the specific end-goal: high-fidelity plain text output (QARI v0.2) or structural document reconstruction with greater training efficiency.

6 Conclusion

In conclusion, this paper presented Qari-OCR, a fine-tuned vision-language model that achieves the strongest performance for Arabic text recognition by leveraging extensive synthetic data and specializing the Qwen2-VL architecture. Our QARI v0.2 model significantly surpasses existing open-source solutions in accurately handling diacritics, diverse fonts, and complex layouts in printed Arabic. Future work will focus on addressing current limitations by enhancing robustness to dense text and embedded graphics, improving numeral recognition, advancing layout analysis for peripheral text, and extending capabilities to Arabic handwriting recognition. These efforts aim to develop Qari-OCR into an even more comprehensive solution for Arabic document understanding.

7 Limitations

Despite the strong performance of Qari-OCR, particularly QARI v0.2, the current study and model possess certain limitations; Firstly, while proficient with dense printed text, the model may encounter difficulties with extremely heavy text layouts where character or line spacing is minimal, potentially leading to recognition errors. Secondly, Qari-OCR's current capabilities are primarily focused on textual content within the main body of documents; it often struggles to accurately recognize and extract text embedded within figures,

charts, or complex graphical elements. Thirdly, the model's performance on historical or non-standard Arabic numeral systems has not been extensively validated and may be suboptimal. Finally, text elements typically found on the periphery of scanned pages, such as book titles on covers, page numbers, or marginalia, are sometimes skipped or inaccurately transcribed, indicating an area for improved contextual awareness and layout analysis.

Acknowledgments

References

I Saleh Al-Sheikh, MASNIZAH Mohd, and L Warlina. 2020. A review of arabic text recognition dataset. *Asia-Pacific J. Inf. Technol. Multimedia*, 9(1):69–81.

Naseem Alrobah and Saleh Albahli. 2022. Arabic handwritten recognition using deep learning: A survey. *Arabian Journal for Science and Engineering*, 47(8):9943–9963.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.

Gagan Bhatia, El Moatez Billah Nagoudi, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2024. Qalam: A multimodal llm for arabic optical character and handwriting recognition. *arXiv preprint arXiv:2407.13559*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. 2025. Ain: The arabic inclusive large multimodal model. *arXiv preprint arXiv:2502.00094*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mistral AI Team. 2025. Mistral ocr: Introducing the world's best document understanding api. https://mistral.ai/news/mistral-ocr. Research, March 6, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Binod Kumar Pattanayak, Anil Kumar Biswal, Suprava Ranjan Laha, Saumendra Pattnaik, Bibhuti Bhusan Dash, and Sudhansu Shekhar Patra. 2023. A novel technique for handwritten text recognition using easy ocr. In *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pages 1115–1119. IEEE.

Joan Puigcerver. 2017. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 67–72. IEEE.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

UNESCO. 2024. World Arabic Language Day. UNESCO Official Website. The Arabic language is a pillar of the cultural diversity of humanity. It is one of the most widely spoken languages in the world, used daily by more than 400 million people.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Mohamed Yousef, Khaled F Hussain, and Usama S Mohammed. 2020. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. *Pattern Recognition*, 108:107482.